

Support for a pragmatic health minister

Cancer patients in Italy are threatening their own survival through faith in a miracle cure. But the government is justified in sanctioning controlled tests of the therapy, even if it lacks a scientific basis.

Astonishingly and scandalously, a frail 85-year-old physician, Luigi Di Bella, has managed to precipitate a crisis in relations between the public, science and government in Italy with his brand of cancer therapy, a cocktail of vitamins and minerals that also includes the drug somatostatin. Dismissed as ineffective by the scientific establishment, the therapy has nevertheless been hailed as a miracle cure by the public.

Early experience with somatostatin in the 1980s and early 1990s indicated that its theoretical promise as an anticancer agent was limited to some rare neuroendocrine tumours. Nonetheless, Di Bella, postulating that somatostatin stimulates the body to rid itself of any type of cancer, has claimed to have cured thousands of patients (see *Nature* 391, 217; 1998). But he has no documentation or publications to back his claims. The health minister Rosy Bindi was therefore right earlier last year to ignore requests from so-called Di Bellists to conduct clinical trials for somatostatin in cancer and to place it on the national list of reimbursable drugs.

But the situation has changed dramatically: this is no longer simply a scientific issue but has become polarized and politicized. Demonstrations of tens of thousands have become a regular event, football supporters wave banners demanding freedom of treatment, and reports of cancer sufferers dying because they were denied the Di Bella treatment dominate newspaper headlines and television debates.

The right-wing opposition party Alleanza Nazionale has formally adopted a policy supporting Di Bella — to be against him is now seen to be left wing — and, more worryingly, the judiciary has backed Di Bella's case against the government, citing an article in the constitution that guarantees to protect the health of individuals. Local courts have ordered individual patients to be treated with somatostatin without charge, and a national court

last month, defying Bindi, ordered the head of the government's pharmaceutical committee to put somatostatin on the national reimbursable drugs list — an order that has since been overturned by a higher court. Italy's constitutional court is now considering the rights of the judiciary to overrule the government on health issues.

In view of this turmoil, Bindi was right when she decided in January to reverse her decision and order clinical trials to be carried out as soon as possible, despite a lack of scientific basis for the therapy. Ten multicentre phase II trials, designed by the country's leading oncologists, began last week, at an estimated cost of IL20 billion (US\$11 million).

The drama of the Di Bella phenomenon may be perceived as peculiarly Italian, but at its core there are parallels with the history of patient power elsewhere. In the United States, patients dying of AIDS successfully persuaded the Food and Drug Administration (FDA) to deviate from its traditional slow drug licensing procedures and make AZT available before clinical trials were completed.

And in the 1980s the FDA was forced to conduct clinical trials on laetrile, a cyanide containing extract of apricot stones of even less therapeutic value than somatostatin. In both cases, concessions by government proved productive in calming the situation; laetrile is a fad of the past, whereas AZT has found its natural level in AIDS therapy.

Some Italian scientists, at least, deplore the sidelining of science as the government, as they see it, bows to emotional pressure in the Di Bella case. They are right to state the scientific case but wrong to oppose Bindi's pragmatic solution. To ignore the emotional element in the public response is to omit a critical factor from the problem and thereby render it insoluble. □

Diplomats need science

A lack of scientific and technical expertise among diplomats is likely to marginalize the diplomatic corps.

In response to recurrent concerns that the US State Department is unable or unwilling to integrate science into US foreign policy, the National Academy of Sciences has been invited to conduct a fresh review of the status of science in the department (see page 427). No doubt the academy will find sizeable flaws in its handling of scientific issues, reflecting the view of science policy experts — expressed in testimony before the House of Representatives Science Committee last week as on previous occasions — that the State Department neglects science. Reforming this situation in an institution that is so large and entrenched is clearly more easily said than done.

Before spending much energy on the attempt, the community should ask itself to what extent the State Department's alleged neglect really matters to research. In one or two respects it does.

Neglect of science and technology at the diplomatic level makes it harder to forge international agreement on scientific collaboration. This weakness has inflicted some damage on various large scientific projects, such as the abandoned Superconducting Super Collider. For the wider global scientific enterprise, the State Department and its equivalents in other countries matter less and less. In science as in commerce, except where intergovernmental agreements are concerned, the process of globalization will increasingly allow international communication to bypass official channels altogether.

But, no less than in the Cold War era, science and technology collaboration can help to build or reinforce good relations between countries. In short, the State Department needs science rather more than science needs the State Department. □



Legal fight looms over patent bid on human/animal chimaeras

[LONDON] One of the biotechnology industry's fiercest critics has fired a shot across its bows by applying for a broad patent on methods for creating 'human/animal chimaeras', a description covering a wide range of experiments in which human cells are fused into an animal embryo, or vice versa.

The patent has been applied for jointly by Jeremy Rifkin, president of the pressure group Foundation on Economic Trends in Washington DC, and author of several books attacking the use of genetic engineering techniques, and Stewart Newman, professor of cell biology and anatomy at New York Medical College, and a long-standing member of the Council for Responsible Genetics.

Rifkin describes it as the "ultimate [patent] prize", given the many potential applications of introducing human cells into animal embryos. Research fields range from testing drugs for teratogenic effects to xenotransplantation.

So far, a combination of scientific difficulties and ethical reluctance has constrained researchers from experimenting with the creation of chimaeric embryos involving human cells. Indeed, in some countries — including Britain and Germany — this is illegal.

Rifkin and Newman have exploited the fact that, in the United States, where no such

law exists, a patent application need not be based on an actual experiment, but can be based merely on the description of a hypothetical experiment, provided the patent office can be persuaded of its credibility.

"There is no need to have actually carried out the experiment, providing the idea meets the standard criteria for patentability," says Pat Coyne, of the Washington-based attorneys Collier, Shannon, Rill and Scott, who have filed the patent. "We are certainly not aware of any 'prior art.'" Coyne also argues that a lack of evidence that others have made similar proposals means that the application meets the claim of 'non-obviousness'.

The two applicants say that if they are granted the patent, they will use it for 'genetic conservancy', to prevent the commercial exploitation of the techniques before there has been a full public discussion of their implications.

If the patent application is rejected by the US Patent and Trademark Office, Rifkin and Stewart say they will take it through the full legal appeals process, including if necessary to the Supreme Court, in order to generate a detailed debate on the extent to which human life is patentable.

Rejection by the court would also have



Mixing and matching: experiments in the 1980s showed that chimaeras like this 'geep' are biologically viable.

important implications for any other patent application on techniques in the same field.

The issue has been highlighted by the fact that the patent applied for on techniques used to produce Dolly the sheep would embrace their use to clone humans (see *Nature* 387, 217; 1997). Under US law, humans cannot be patented, because this would contravene the Thirteenth Amendment to the constitution, outlawing human slavery.

But in 1980 the US Supreme Court handed down a landmark decision, based on an application from Ananda Chakrabarty covering a microorganism that had been modified to help clean up oil spills. It ruled that there was no bar to patents on living organisms, which can be considered as "compositions of matter" (see *Nature* 285, 528; 1980).

Although the same principle applies in Europe, illustrated for example by the decision to grant a patent on Harvard University's 'oncomouse' (see *Nature* 353, 589; 1991), a new directive on biotechnology patents explicitly rules out human/animal chimaeras.

But in the United States, even industry lawyers admit that there is no legal consensus on the boundary between the types of life-form that can and cannot be patented. "Where the crossover point lies is certainly a grey area," says David Mickel, legal counsel to the Biotechnology Industry Organization in Washington DC.

The application describes three separate techniques for creating biologically viable chimaeras between species. The first, which involves mixing cells from the early embryos

Africa defends rights to indigenous knowledge

[LONDON] All countries belonging to the Organization of African Unity (OAU) have agreed to consider refusing to recognize any patent on a drug made from natural products found in Africa — unless it acknowledges the 'ownership' and contribution of the relevant community to the new product.

A model bill produced by an OAU committee states that ownership of new compounds should rest with indigenous local communities for "all times and in perpetuity".

It calls for states to develop laws guaranteeing such ownership. It also calls on collectors of natural products to share information with governments on "all discoveries" from research and development.

The draft bill was drawn up to harmonize African legislation on 'bioprospecting' by multinationals.

It was done partly to challenge the Trade Related Intellectual Property Rights Agreement (TRIPs) of the World Trade Organization (WTO), and partly to clamp down on the smuggling of medicinal plants.

The wording was finalized at a meeting of the OAU's scientific, technical and research commission in Addis Ababa, Ethiopia, last week. It will now be circulated among the OAU's 53 member states for comment before being presented as model legislation for African states.

The OAU believes that the TRIPs agreement violates the United Nations

Biodiversity Convention, which makes the "approval and involvement" of indigenous peoples a condition of developing a product based on a natural compound. According to the OAU, the TRIPs agreement makes no such provision.

Johnson Ekpere, the commission's executive secretary, says African countries want to bring TRIPs into line with the biodiversity convention.

But the draft will be controversial.

Some African states, such as Nigeria and Tanzania, are keen to attract overseas bioprospecting partners. They may be unwilling to pass legislation that could scare them away. **Ehsan Masood**

of two or more different animal species, was used to create the sheep-goat chimaera in Britain in the 1980s (see *Nature* 307, 634–636 and 637–638; 1984).

In the second technique, which is widely used to introduce targeted mutations into mice, undifferentiated embryonic stem (ES) cells from the inner cell mass of preimplantation embryos are combined with normal preimplantation embryos.

The third technique combines so-called 'early passage' ES cells, those that have undergone only a few cell divisions, with embryos that have been altered to prevent them advancing beyond an early stage of development, but which will still allow the implanted ES cells to differentiate normally to form viable embryos.

In their patent application, Rifkin and Stewart say they can envisage ways in which one or more of these techniques could, in principle, be used to produce chimaeric embryos containing both human and animal cells that would have important applications for biomedical research. In particular, they identify:

- mouse/human embryos that could be used in developmental biology studies, particularly in investigating how cells from the two sources cooperate to form an embryo;
- baboon/human or chimpanzee/human embryos, of particular value to pharmaceutical and chemical companies for studying the teratogenic effects of new compounds;
- chimpanzee/human chimaeras which could in principle — and provided that developmental incompatibilities between blastomeres or ES cells were overcome and the embryo brought to term — "provide

ideal test systems for the cardiovascular effects of whole animal stress"; and

- chimpanzee/human and pig/human chimaeras, used for example as a source of hearts for transplantation to cardiac patients.

The breadth of the application, which is also intended to cover "the modifications and variations" of the techniques described, is already leading some patent lawyers to argue that, at the very least, its scope will be severely restricted by patent examiners before it is granted.

Its audacity leads others to dismiss it as little more than a joke. "This seems to be just a publicity stunt," says one British patent attorney (a related application is to be filed with the European Patent Office). "Chimaeras raise all kinds of emotive issues that go back to Greek mythology; I am rather bemused by the whole thing."

But Rifkin, who says there has been no serious public discussion of the patenting of life-forms in the United States since the Chakrabarty decision — which was itself approved by the Supreme Court by only a 5:4 majority — insists this is a serious attempt to stir up debate in political and legal circles.

"By having control of this patent, we will be able to provide countries around the world with the opportunity to debate the full implications," he says. "We think there should be a period of time for robust debate to establish the necessary guidelines."

Newman, too, is keen to have this discussion before the techniques become widely used. "No one has yet announced that they are going to do this type of thing," he says. "But it is clearly the type of thing that people would find useful, and I would far rather that



Rifkin: seeking a "genetic conservancy"

we test the process and be in control than let it roll along with no public debate."

Jonathan Marks, who teaches biology and anthropology at the University of California, Berkeley, and includes discussion of the moral implications of crossing chimpanzees and humans in his lectures, says that "if this is what it takes to encourage geneticists to think more about humanistic issues, I am all for it", although he admits that he "would rather see it done more painlessly".

The process of dealing with initial comments from the US Patent Office, and if necessary contesting legal decisions through various stages, is likely to take years. During this time, as a patent applicant, Rifkin will have the legal standing to comment on similar applications made by others.

Given this prospect, many in the biotechnology industry are likely to see the application and the publicity campaign it is intended to stimulate as frivolous and irritating at best, and potentially disruptive at worst.

But even some geneticists who disagree with the anti-genetic engineering thrust of many of Rifkin's arguments accept the legitimacy of the issues he raises. "Setting aside the biological arguments, this is very much a mix of a legal and philosophical discussion," says David Porteous, of the UK Medical Research Council's Human Genetics Unit in Edinburgh. "It would certainly be a useful extension of the debate."

David Dickson

Neuroscientist accused of misconduct turns on his accusers

[SAN DIEGO] A federal hearing whose outcome could have broad implications for US academic institutions began this week in Houston, Texas, about charges that a neuroscientist committed scientific misconduct.

The case has been watched closely by academic institutions, because it involves one of the most aggressive defences in the nascent history of formal misconduct inquiries: a legal attack on the accusers.

Kimón J. Angelides, who is 46, formerly of Baylor College of Medicine and now at the University of Durham in the United Kingdom, is appealing against a proposed five-year debarment by the National Institutes of Health (NIH) for falsifying and fabricating research data in grant proposals and five published articles (see *Nature* 383, 107; 1996).

His appeal is being heard by a three-person panel during two-week proceedings, with lawyers for the government and Angelides taking testimony from and cross-

examining about 25 witnesses, including Angelides himself.

The panel is to decide in June whether Angelides committed scientific misconduct — which both Baylor and the NIH's Office of Research Integrity (ORI) have previously ruled he did. Angelides was fired by Baylor in 1995 after the university concluded a lengthy investigation of the alleged misconduct between 1988 and 1992.

Angelides is suing Baylor and its officials in state court for wrongful termination, defamation and other offences. Angelides denies any misconduct, blaming any irregularities on other scientists who had worked in his laboratory studying the sodium channels of nerve cells. Angelides' civil lawsuit and misconduct allegations have been the subject of varied litigation.

Baylor unsuccessfully attempted to remove the lawsuit to federal court last year, pleading immunity from a lawsuit because it was following NIH guidelines on grant-monitoring. The US Attorney's Office in

Houston conducted an inquiry into Angelides, but closed it last summer without filing any criminal charges. Now the state trial on Angelides' lawsuit is set for October.

The outcome of the federal appeal hearing is likely to have a major impact. If the government's and Baylor's findings of misconduct are upheld, Angelides may have a difficult time winning his lawsuit. If the findings of misconduct are overturned, Baylor could find huge costs awarded against it by a jury hearing the lawsuit.

A government loss in the case and damage award against Baylor could embolden other accused scientists to follow a similar litigation course, thereby undermining government fraud-prevention efforts and sending a chilling message to academic institutions.

Angelides' co-authors for the five articles in question — most of whom are from Yale University — have sought to retract all five, but only two retractions have been published so far.

Rex Dalton

Space agency adopts user-led strategy...

[PARIS] The ruling council of the European Space Agency (ESA) last week adopted a new strategy that marks a significant shift in philosophy. Its top-down, bureaucratic organization is becoming more streamlined, geared to the needs of scientists, industry, government and other 'users'.

The strategic plan by the ruling council, made up of the heads of space agencies, will be tested later this year, when ministers from member states meet to decide whether to endorse the new directions, and to agree on funding for the proposed programmes. Sources in several states predict that the strategy will attract a broad consensus from space ministers, although funding for the science programme remains contentious.

The ministerial meeting was expected in June, but may be delayed because of the German general election in the autumn and because smaller member states are unhappy with the current lack of a firm commitment on science spending (see below).

The strategy bears the stamp of Antonio Rodotà, an industrialist who was appointed director general of ESA last year with the remit of reforming the agency. It has already

trimmed its 3,500 staff by a quarter over the past three years. The proposed changes reflect the general shift by member states, already evident from the last ministerial meeting in 1995 (see *Nature* 377, 667; 1995), away from prestige projects such as an independent manned space programme, towards more commercial and social goals.

UK science minister John Battle has strongly endorsed the new direction, in a marked departure from the hostility that has characterized Britain's attitude to ESA over much of the past decade.

"I've been stunned at the degree to which ESA has adopted what the UK and others have been saying," says one official at the British National Space Centre, cautioning that much work remains to be done in putting the new philosophy into practice.

The flagship of the approach is a proposed comprehensive programme in Earth observation. Its structure is based on ESA's science programmes, with participating member states required to approve overall funding for five years — around ECU 450 million annually — and ESA deciding how money should be allocated to proposals by users.



Making space pay: satellite images, such as this from ERS2, can help control pollution.

David Southwood — professor of physics at Imperial College, London, and currently on secondment to ESA to help develop the programme — calls this a "radically new way of doing business" and says the five-year funding will provide a stability that is lacking in the current approach, in which missions are approved one at a time. This has resulted in a lack of forward planning, says Southwood, who points out that ESA has no plans for observation satellites beyond Envisat, scheduled for launch in 1999. Budgetary stability should also facilitate collaboration with US and Japanese space agencies on long-term ventures. Global programmes to study major scientific challenges are already under discussion, he adds.

The programme will focus on jointly developing basic science and applications, with the launch of several science and technology demonstrators called Earth Explorers. The first four priority areas in this programme, headed by ESA's head of science Roger Bonnet, are atmospheric dynamics, the Earth's radiation budget, a gravity mission and a land surface hyperspectral. A series of optional Earthwatch missions will be conducted to develop commercial and public service applications.

The main challenge facing ESA is to make a smooth transition from technology demonstration to operational use, says Southwood, pointing out that Europe often lacks the mechanisms for using space data, such as ocean studies. Many feel that ESA should focus on research and development, while working closely with industry and the European Commission to transfer technologies when these are ready for exploitation.

ESA and the commission intend to sign an agreement outlining mechanisms to introduce Earth observation into agriculture, maritime and other sectors. Another goal will be to reduce duplication between national and European efforts, says Southwood. "Having got the programme rolling, a vital need is to work out how we make it a European programme."

Declan Butler

...but leaves question mark over science funding

[PARIS] Uncertainty is hanging over Europe's joint science space programmes. Members of the European Space Agency (ESA) are expected to demand cuts when ministers meet later this year to endorse a new space strategy.

ESA's ruling council reaffirmed support for the programme last week. But agreement on funding for Horizons 2000, ESA's dedicated basic science programme, over the next five years was conspicuously absent from its strategy.

"There is an issue in governments' minds about how important basic science is, how much it deserves," says one UK official.

Ministers seem unlikely to lift a cap on science spending, imposed two years ago at the United Kingdom's demand, that has cut the science budget by 9 per cent. Sweden and France are believed to share UK opposition to lifting this cap, while Germany and Spain

are said to want bigger cuts.

Roger Bonnet, ESA head of science, says budgets are already strained to breaking point and any new cuts would seriously disrupt a series of science missions planned for the next decade. ESA's Science Programme Committee has already warned that Mars Express, a ECU150 million (US\$163 million) mission to Mars, will probably be cancelled unless the cap is lifted.

Bonnet has tried to reduce costs by delegating management to industry and using technologies from existing missions. He has oriented the agency's ECU300 million for 'medium-sized missions' towards smaller, flexible ones, including technology-testing missions (see *Nature* 385, 380; 1997). But such measures increase the risk, he adds.

One UK official argues that the cap has promoted cost-effectiveness. "There is no way ESA would have in the past proposed a mission

to Mars with a price tag as low as ECU150 million," he says. "Bonnet has gone three-fifths of the way, but still has some distance to go."

But David Southwood, professor of physics at Imperial College, London, and currently on secondment to ESA, says there is little scope for further economies and warns against the risk of "service degradation".

There are also fears that the total cost of programmes in the proposed ESA strategy (see above) is likely to exceed the annual budget of ECU2.5 billion. Leaving aside the prospect of cost overruns, ESA's contribution to the international space station will increase from ECU300 million now to ECU500 million in 2000, while the proposals contain plans to develop more powerful launchers based on the new Ariane 5 rocket. Bonnet warns that competing pressures on the budget mean science risks paying for any shortfall.

D. B.

Asian states take 'first step' on acid rain

[TOKYO] Concern about the problem of acid rain in East Asia has led nine countries, including Japan, South Korea, Indonesia, Thailand, Malaysia and Russia, to discuss setting up a regional monitoring network.

The proposal was the subject of a meeting in Yokohama, Japan, last month, which had been called for by the government of Japan. In addition to delegates from the nine countries involved, observers from China, the World Meteorological Organization and the United Nations also took part.

No-one is expecting an international emission control agreement on sulphur dioxide in East Asia to emerge easily. But officials at Japan's Environmental Agency argue that cooperative research and monitoring could pave the way towards concerted action.

Total sulphur emissions from 20 Asian countries have been estimated at 34 million tonnes in 1990, with north Asian countries accounting for more than three-quarters. According to calculations by the World Bank before last year's financial crisis, emissions could rise to 78 million tonnes by 2010 if no firm action is taken. In Europe, sulphur emissions are expected to decrease from 38 million to 14 million tonnes over this period as a result of stringent controls.

A consensus document agreed by delegates at the meeting calls for a two-year preparatory phase of the monitoring network, starting this month. During this period, a few monitoring sites will be set up in each country. Financing throughout the preparatory phase will be provided by Japan, which will also host the network's secretariat and data analysis centre.

But the consensus document says that further discussions are needed on several

issues. These include the role of the scientific advisory board to be attached to the network, the implementation of quality control protocols, the location of the permanent network centre and secretariat, and the possibility of setting up branches of the network centre in other countries. The final shape of the network will be discussed at a second inter-governmental meeting scheduled for 2000, the year in which the network is expected to start full operation.

Delegates to the meeting said they welcomed the Japanese initiative. But a spokesperson for the Japanese foreign ministry admitted at a press conference after the meeting that there remains suspicion on the part of other East Asian countries as to a possible political motivation for the network.

Such suspicion is understandable, as Japan is keen to demonstrate that an increasing percentage of the acid rain falling on the country originates in the industrial districts of north China and Korea. A Japanese proposal to provide a permanent home for both the network's centre and its secretariat appears to have been vigorously opposed.

Further suspicion was generated because the intergovernmental meeting was convened only after the design of the network had been drawn up in informal expert meetings organized by the Japanese Environmental Agency. At these meetings, scientists from Japan outnumbered those from other countries.

Delegates from other Asian countries argued that the network design should be reviewed by the scientific advisory body, on which scientists from all participating countries are equally represented, says Seok-Young Choi, director of the environment and

science division at the South Korean Ministry of Foreign Affairs and Trade. Delegates emphasized "the need to enhance full transparency in organization and operation of the network", says Choi.

Alexandre Soudine, senior scientific officer at the environmental division of the World Meteorological Organization, who attended the meeting, says that "a first step has been made" towards addressing the acid rain problem. He added that his agency "supports the Japanese initiative".

The organization has released an assessment report on the global state of acidic deposition last year. Soudine points out that reliable data on acidic deposition are not yet available for many parts of the world, including East Asia. He argues that, to assess the problem in East Asia, considerable modelling will be needed in addition to monitoring.

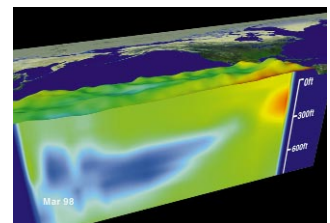
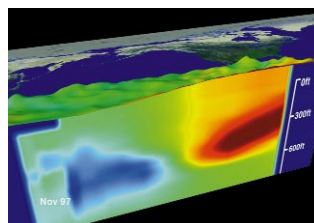
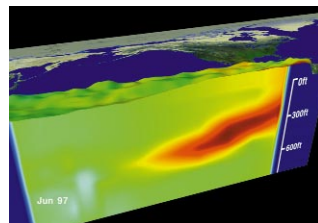
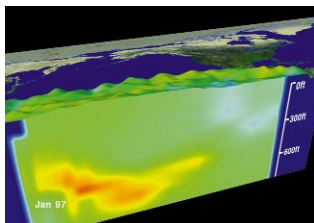
The European Monitoring and Assessment Programme, which has served as a model for the Japanese initiative, gathers data from more than a hundred monitoring sites all over Europe. In contrast, the East Asia network will initially have few monitoring sites.

Initial modelling has been carried out by the International Institute for Applied System Analysis in Vienna, and by Japanese research institutes. In a five year-project funded by the World Bank, the Vienna institute has adapted its RAINS model of acidification to the Asian situation.

The Asia version of the RAINS model is commercially available and has been distributed widely in Asia. But critics say that it provides only a very broad assessment of the acidic deposition problem in the region, and lacks sufficient detail to be of significant value.

Robert Triendl

El Niño in the frame for Pacific temperature and height increases



[LONDON] Three forms of data have been used to provide a dramatic three-dimensional visualization of increased sea level height and raised sea temperatures in the eastern Pacific linked to the El Niño event (above). Data from three separate instruments, including the US National Aeronautics and Space Administration's (NASA)'s Topex radar altimetry satellite, released last week, show a 34-cm rise in the height of the eastern Pacific Ocean between January 1997 and March this year. Sea

surface temperature increased by 5.4 °C during the same period.

The height of the sea is indicated by the bumps in the images. Sea surface temperature is represented by the colour, with red being 10 °C above normal, and blue 10 °C below normal. The large area of blue in the final frame indicates colder than normal water being left behind as warmer water moves further eastwards.

The winter El Niño forecast of warmer waters moving eastwards across the Pacific

Ocean was issued last summer. Data from the instruments show that the 1997–98 El Niño arrived earlier than the last comparable event in the winter of 1982–83. But projections of drier than normal weather in parts of Asia have not quite materialized (see *Nature* 388, 108; 1997). India, for example, experienced a near-normal monsoon.

The frames, as well as a computer simulation of the movement of water, can be viewed on the NASA El Niño Web site, http://nsipp.gsfc.nasa.gov/enso/nino_update.

US foreign policy under fire over strategy on science

[WASHINGTON] The US State Department, criticized in recent years for paying insufficient attention to science and technology, has turned to the National Academy of Sciences for advice on how it should best integrate science policy with foreign policy.

In a letter sent to the academy last month, legal counsel to the department, Wendy Sherman, requested that it conduct a study of the department's science programme. The academy's governing board will consider the request later this month.

Sherman's letter conceded that the State Department "may not be doing as much in the science, technology and health areas as we can". One reason, says Melinda Kimble, Acting Assistant Secretary for Oceans and International Environmental and Scientific Affairs, is a belt-tightening throughout the department that occurred in the mid-1990s.

Funding for science, technology and health is only now beginning to make a modest comeback under Secretary of State Madeleine Albright. Historically, scientific agreements with other countries were associated with Cold War strategy, says Kimble, and became lower priority when the Cold War ended.

Only recently, she says, have senior officials at the department, such as Timothy Wirth, until recently the Under Secretary for Global Affairs, and Stuart Eizenstat, Under Secretary for Economic and Business Affairs, begun promoting science as integral to US environmental and trade strategy.

A 1992 report by the Carnegie Commission on Science, Technology and Government called for "deep-seated reforms" in the way US international science policy is conducted. These included the creation of the post of Counselor for Science and Technology at the department, and a boost to funds and staff for science, technology and health.

But James Watkins, president of the Consortium for Oceanographic Research and Education, says the reforms have not happened, and "if anything, the situation has worsened". Watkins and other witnesses at a congressional hearing last week said that US international science policy remains disjointed. They argued that agencies such as the energy and defence departments and the space agency NASA are left to negotiate their own deals, with little coordination from the State Department or anyone else. The hearing was part of the House of Representatives Science Committee's review of national science policy (see *Nature* 386, 100; 1997).

Watkins, a former energy secretary under George Bush, complained that "mega-projects, such as the Superconducting Super

Collider and the International Thermonuclear Experimental Reactor, have suffered from the department's lack of involvement. Climate research and oceanographic studies will be similarly impaired, he says, unless there is "radical surgery on today's ineffective system".

His suggestions include placing scientists in US embassies and for congressional science and foreign relations committees to hold meetings with Albright to emphasize the importance of science to foreign relations.

But Thomas Ratchford, a science policy analyst at a research centre at George Mason University in Arlington, Virginia, suggested looking outside the State Department for answers. Efforts at reform "have in general failed," he said. "It is time to quit trying to fix the system directly through State and to enable the technical agencies to partially fix the system."

Ratchford said other countries often fill diplomatic science posts with people borrowed from technical agencies — a practice the United States may want to try, as career foreign service officers tend to be generalists without scientific expertise. The National Science Foundation would be the logical agency to supply or screen these scientific staff, he said. Kimble hopes the academy study will consider how the department can work with other agencies to help achieve its own objectives.

The witnesses disagreed about how effective international scientific agreements have been. About 3,000 US projects funded in 1995 had some international component, according to a study completed last year by the Rand Corporation at the request of the White House science office. But fewer than 10 per cent were coordinated through the State Department, says Caroline Wagner, a senior policy analyst at Rand. The smaller agency-to-agency agreements tended to be the most effective, she said, and have had a "substantial benefit to the United States".

Ratchford, however, criticized the kinds of cooperative agreements often signed during state visits. "In general these agreements have not been very productive scientifically or diplomatically," he said. "There is the diplomatic and public relations benefit of the initial signing ceremony, but the scientific and technological results have often been nil or close to it."

One reason, Ratchford said, is that "it is difficult for diplomats and White House staffers with little experience in research to construct the framework for a viable research effort in the short time available for planning a presidential trip". **Tony Reichhardt**

Scientists defy their ethics codes and take gifts from industry

[WASHINGTON] Life scientists are eagerly accepting gifts from industry even when there are strings attached, according to a US journal. The materials, equipment and trips received often defy universities' ethics rules, says a study in this week's *Journal of the American Medical Association (JAMA)*.

Almost half (43 per cent) of more than 2,000 researchers at 50 top US research universities surveyed for the study (see *JAMA* 279, 995; 1998) had accepted a gift in the past three years. The most common gifts were biomaterials (24 per cent), money (15 per cent), research equipment and trips to meetings (11 per cent each).

In many cases, however, researchers who accepted the gifts felt that fairly onerous conditions were attached to them. One-third of gift donors required pre-publication review of any resultant research, and one in five demanded ownership of all patentable results. These demands were most frequently made by donors of biomaterials, such as assays and cell lines.

Eric Campbell, a sociologist at the Health Policy Research and Development Unit at the Massachusetts General Hospital at Boston who conducted the study with a grant from the National Institutes of Health, believes both sets of conditions are cause for concern.

Campbell says previous research of his suggests that industry can ask for publication delays of as long as six months to allow for review. He also points out that most universities have explicit policies on intellectual property rights that do not allow investigators to surrender patent rights.

"Gifts from industry to life scientists are a common and important form of academic-industrial research relationship," his study concludes. "At times it may be prudent for faculty members to 'look at a gift horse in the mouth'."

Campbell didn't try to put a cash value on the gifts, saying that the value of such items as biomaterials can be difficult to assess. But two-thirds of recipients regarded the gifts as important to their work.

The study found that male researchers were more likely to get gifts than females, and senior faculty more likely to get them than junior faculty. Recipients were found to be significantly more productive than non-recipients — even when these gender and status differences were factored out.

In an accompanying editorial in *JAMA*, Lisa Bero of the Institute for Health Policy Studies at the University of California at San Francisco argues that the study makes "a compelling argument" for fresh guidelines on the acceptance of gifts. **Colin Macilwain**

Ageing population of voters backs Alzheimer's funding

[WASHINGTON] Groups intent on securing increases in US research funding for particular diseases have recently intensified campaigns in Washington DC in the hope of substantially increasing their shares of the generous 1999 budget increase proposed for the National Institutes of Health (NIH).

Many have been using well-known personalities to present their case, and are



Actress Piper Laurie at last week's hearing.

countering warnings from the scientific community that such lobbying distorts research priorities with appeals both to scientific opportunity and to the economic incentives to target diseases that will affect increasingly large numbers of an ageing population.

The Alzheimer's Association, for example, wants an extra \$100 million — a 28.6 per cent increase — in funding for the disease. This was the subject of a hearing convened by Senator Arlen Specter (Republican, Pennsylvania) last week which featured, among others, the actress Piper Laurie. Specter chairs the subcommittee that writes the bill that funds NIH.

Similarly, advocates for people with Parkinson's disease — including former heavyweight boxer Muhammad Ali, who suffers from the disease — made their case to the health and environment subcommittee of the House of Representatives Commerce Committee, which writes occasional bills setting the broad direction for NIH.

They insisted that doubling the funding for research into Parkinson's disease is essential for adequately exploiting research opportunities. They also argued that Congress is legally bound to produce the funds under a law passed last year calling for \$100 million in new Parkinson's money over each of the next three years, as it had not provided the money (see *Nature* 389, 112; 1997).

The flurry of lobbying comes as Congress sets about drafting spending bills that promise to give the NIH an increase of at least 8.4 per cent, to \$14.8 billion, and perhaps even more. It also comes on the heels of strong objections from NIH directors to congressional interference in research funding.

The directors told a panel convened by the Institute of Medicine last month that efforts by Congress to "earmark" funds for specific diseases "deform" science and, when perma-

nent, are even "lethal" to the research enterprise (see *Nature* 392, 116; 1998).

Such pleas, however, do not always prevail with politicians who may draw on personal experience of family or friends crippled by diseases. For instance, Paul Wellstone (Democrat, Minnesota), a leading backer of increased Parkinson's funding, saw both his parents suffer with the disease.

And Newt Gingrich, the Speaker of the House of Representatives, whose mother-in-law has diabetes, wrote a \$150 million increase in spending on juvenile diabetes research into the balanced budget law last year (see *Nature* 388, 617; 1997).

Richard Hodes, the director of the National Institute on Aging, which funds about three-quarters of NIH Alzheimer's research, says the new money could "very definitely" be spent on first-rate science, and that the ageing of the US population presents a "time imperative" for accelerating the research. Steven Hyman, director of the National Institute of Mental Health, which plans to spend \$28 million on Alzheimer's in 1999, says the field has "immense opportunity".

But both men argue against the Alzheimer's Association's attempts to increase research spending by 28.6 per cent. (The budget that President Bill Clinton sent to the Congress in February would increase it by 7.4 per cent, to \$375 million.)

Hodes says that "scientific planning and scientific resource setting", not Congress, should determine how the NIH spends its money. And Hyman argues that if Congress were to allocate an extra \$100 million for Alzheimer's without increasing the entire NIH budget by the same amount, it would force the NIH to raid accounts for other diseases, and "might actually draw money from better to less worthy science".

But in a country where 14 million people are expected to succumb to Alzheimer's when the baby boomers reach 65, politicians are hard pressed to resist calls to target a costly disease. At last week's hearing, Specter said the Alzheimer's Association's goal of \$100 million in new funding this year is "laudable". And Tom Harkin (Democrat, Iowa), the senior Democrat on the subcommittee, urged advocates to put "maximum" pressure on Congress to vote for the increased Alzheimer's spending.

At the House of Representatives, the reaction from John Porter (Republican, Illinois), the chairman of the subcommittee that funds the NIH, was less encouraging. Porter said that, where the NIH was concerned, political judgement was no substitute for scientific judgement.

Meredith Wadman

Coalition to pursue ethnic concerns over gene research

[COLLEGE PARK, MARYLAND] The concerns of ethnic minorities about genetics research and related policies are to be addressed by a coalition of US academics, clinicians and community advocates. The coalition, which argues that the interests of a substantial proportion of the US population are frequently overlooked by genome researchers, has promised to provide a focus for these concerns.

The multicultural task force was set up last week at the end of a two-day symposium on genetics and ethnic minorities sponsored by Howard University in Washington DC and the Sinai Hospital in Baltimore, Maryland. The organizer, Ilana Mittman, who is director of the genetic counselling programme at Howard, a historically black college, said at the symposium: "The interests of communities of colour must be tied to the design and implementation of genetic policy."

Mittman said that the new group plans to link up with the American Association for the Advancement of Science to organize legislative hearings. The group would press Congress to link funding for genome research to performance in this area. Informed consent, institutional review boards and possible bias in research design, scientific review and funding would all come under scrutiny, she said.

Patricia King, law professor at Georgetown University in Washington DC, said: "The conceptual framework doesn't work any more." She argued that existing structures, such as institutional review boards, broad-based recruitment and informed consent, had each failed to provide sufficient protection for minorities. Researchers needed to find ways of ensuring that those who contribute to research are also able to benefit.

Genome-project administrators at the meeting acknowledged there is work to be done. "I'm here to listen and learn; I would be first to say we're not doing enough," said Dan Drell, who oversees research in the ethical, legal and social implications (ELSI) aspects of the human genome programme for the Department of Energy.

Participants argued that definitions and initiatives developed solely from the perspective of white Americans risked diverting attention from some urgent needs of non-whites. For example, they complained about the decision to offer cystic fibrosis screening to all pregnant women, pointing out that African-Americans, Asian-Americans and Latinos are very unlikely to be carriers.

There was also criticism of researchers for lumping together ethnic groups into broad categories such as Asian-Americans and Hispanics, masking health problems particular to certain subgroups.

Sally Lehrman

Tokyo's plan for new campus under threat

[TOKYO] An ambitious plan by Tokyo University to build an 89-acre campus, 20 miles from the city centre, is hanging in the balance.

The university has run into difficulties in raising funds to buy the land in Kashiwa, Chiba prefecture. Other problems include a lack of enthusiasm among the university's existing departments for the creation of a new 'centre of excellence' at Kashiwa, and concern about the upper age limit of 35 for recruits to 14 planned new research groups.

These groups form part of the university's '21st Century Project', intended to exploit promising new areas of science. Research groups in areas such as materials science, energy, communications, life sciences and environmental science are planned to be set up by April next year.

But this will happen only if the financial prospects improve. In 1995, the university bought 29 acres of the land using funds allocated by the Ministry of Education, Science, Sports and Culture (Monbusho) as part of the supplementary budget. The university intended to sell a site belonging to the Institute of Solid State Physics (ISSP) in Roppongi, central Tokyo, and use the proceeds to buy the remaining 60 acres of land in Kashiwa.

But a recent collapse in Tokyo land prices has reduced the value of the Roppongi campus too far to cover the price of the Kashiwa campus. The university recently failed to persuade Monbusho to purchase the land.

According to Tokyo University, the new research groups, which will start operating

next year, will initially be based at the main campus in Hongo, central Tokyo, but will transfer to Kashiwa later. Some researchers now feel that this is unlikely to happen.

"We have the feeling that the new research groups will stay permanently at the main campus in Hongo," says one researcher in the department of physics.

Others are more optimistic. Shuntaro Watanabe, a professor in advanced spectroscopy at ISSP, suggests that, rather than the Roppongi campus, the university might sell the site of the Institute of Cosmic Ray Research in Tanashi, on Tokyo's outskirts. He also predicts that the university is likely to be granted extra funds from the government's ¥16,000 billion (US\$124 billion) economic stimulus package, unveiled last week to help boost the sluggish economy.

The scheme to create a new centre of excellence for postgraduate research in the suburbs was launched by Hiroyuki Yoshikawa, the former president of Tokyo University. But individual departments have since begun trying to build their own centres of excellence at the existing campuses.

Meanwhile there is concern that the new campus is eating up money that would otherwise be spent on repairing old, dilapidated departmental buildings. Funds are also needed to complete a building for the faculty of science on the main campus (see *Nature* 355, 99; 1992): five years behind schedule, it already houses several departments.

There have also been criticisms of the

university's management over recruitment for the impending research programme. Many scientists were disappointed to find out that advertisements carried strict age restrictions — new assistant professors, for example, could be no older than 35.

Some researchers feel this restriction, intended to bring new blood into the university, is too strict. One professor at the university's Research Centre for Advanced Science and Technology (RCAST), argues that it violates scientists' right to equal opportunities. He says it virtually rules out everyone who has carried out research abroad or in industry, or obtained an MD.

"Although the university is outwardly promoting a policy to appoint 'bright young things', it is evident that they are creating a bureaucratic fast-track system for young, but not necessarily talented, researchers within the university," he says.

Some researchers even fear that the new assistant professors have already been selected by staff who lead powerful factions within the university. They are concerned that this practice, apparently not unusual in the university, contradicts its aim of producing internationally competitive researchers.

Protests by angry scientists are said to have forced the university committee to remove the age restriction, while RCAST, a comparatively forward-thinking institute within Tokyo University, is considering introducing an equal opportunity policy in the near future.

Asako Saegusa

Russian students warn of demonstrations over spending cuts

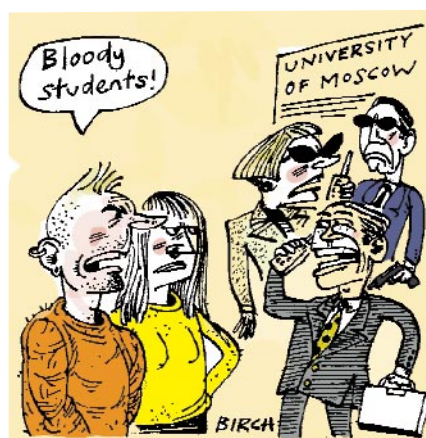
[MOSCOW] Students have launched a nationwide protest against proposals for university reform by Russia's minister of education, Alexander Tikhonov. They say the reforms could lead to a steep fall in the number of graduates and a significant increase in students' costs.

The Russian Association of University Students Trade Unions has sent an official note of protest to Tikhonov, President Boris Yeltsin and Gennagy Seleznev, the speaker of the Duma (the lower chamber of the Russian parliament).

The association has 2.6 million members, 70 per cent of Russia's student population, and plans to hold a nationwide protest that is likely to affect public opinion.

Last year the streets of the major university towns, including Moscow, St Petersburg, Nizhny Novgorod and Tula, saw local demonstrations, but these were little more than the traditional protests against the usual difficulties faced by students.

This time, the students argue that the reforms threaten to undermine the whole



higher education system. Their action could paralyse cities and block federal highways.

"We will try to avoid extremism, but cannot guarantee order," said Mikhail Mirsky, deputy chairman of the student association. "Tikhonov has created a scheme that could lead to the total annihilation of Russian universities. He will only be helping students who belong to rich families. The

number of students paid for out of the state budget is to be cut by 10 per cent."

Mirsky points out that these changes are taking place as the price of accommodation is soaring, and students must for the first time pay for textbooks. "We predict that within three to four years the number of students will have fallen by 40 to 50 per cent, at a time when the government increasingly needs graduates," he says.

Tikhonov, son of the former Soviet prime minister Nikolai Tikhonov, has said he intends to solve the growing problems of Russian higher education by "searching for resources outside the state budget".

For example, he is suggesting that universities should be able to lease buildings to private companies. But he also wants a reduction in the number of universities, the length of courses — from six to four years — and the ratio of lecturers to students. The student association complains that such moves will inevitably lead to widespread corruption and to campuses being occupied by businessmen.

Carl Levitin

Europe's millennium Moon mission fails to get off the ground

[PARIS] Proposals for a US\$200 million Moon mission to celebrate the millennium that were announced with great fanfare by the European Space Agency on 5 March were quietly abandoned at a meeting of the agency's ruling council last week (see page 425). The proposals called for the launch in 2000 of a small orbiter, Lunarsat, to scout landing sites at the Moon's south pole, followed the year after by the sending of a lander to begin setting up a 'robotic village'.

The proposed mission has been criticized by some as a stunt of limited scientific interest (see *Nature* 390, 8; 1997). Much of the funding was intended to come from industry and commercial sponsors. Some observers argue that this finance was likely to have materialized had the agency's council agreed to pump-prime the project with a small sum of money.

But the proposal died a quick death after a brief discussion by the council, which is said to have feared committing itself to a project whose costs might spiral if other sponsors failed to materialize.

Hands off our intellectual property, say researchers

[MUNICH] ALLEA, the association of 50 European academies of science, is calling for better protection of European researchers' intellectual property rights. It claims that the interests of the scientific community in Europe have been largely ignored in drafting recent directives on such rights, such as the European Union biotechnology directive and the database directive.

At the association's third general assembly, held in Munich last week, ALLEA members agreed to make a formal request to the European Parliament and the union's council of ministers to introduce into the harmonized patent law a 'grace period' allowing researchers to apply for patents for a defined period after publication. This 'grace period' should address the problem faced by European participants in the Human Genome Project, as the requirement to release sequence data immediately on to the World Wide Web eliminates the possibility of protection in Europe.

Russian academy opens door to more women

[MOSCOW] A two-day general meeting of the Russian Academy of Sciences awarded a gold medal to the academy's president, Yuri Osipov, for his achievements in mathematics. It also welcomed the news that the latest elections brought nine women

members — an unprecedented number — into the academy.

The meeting debated the implications of a new agreement with the state property committee, allowing the academy to keep for its own use the money it earns by leasing property given to it by the state. In a separate move, the academy decided to take responsibility for managing — and reaping the financial rewards of — the sale of rights to inventions produced in its laboratories.

French demand DNA database to fight crime

[PARIS] Pressure is growing in France for the creation of a national database of DNA profiles from convicted criminals, following the arrest last week in Paris of a suspected serial killer, accused of at least five murders and rapes. The suspect, Guy Georges, was identified after Gilbert Thiel, the judge leading the investigation, found a match between DNA at the scene of the crimes and DNA taken earlier from the suspect, who has a string of past serious criminal convictions.

Thiel, who personally searched laboratory records, said that the lack of a national database had "certainly been an obstacle" and had delayed the arrest. Plans for a national database, limited to those convicted of serious crimes, are contained in a bill on sexual delinquency submitted to the National Assembly last year. But an amendment forbidding conservation of DNA samples has resulted in confusion about the bill's scope.

Oil boss honoured for backing sustainability

[ST LOUIS] John Browne, chief executive of the British Petroleum company, has been honoured by the Missouri Botanical Garden for his efforts on behalf of sustainable development. In a ceremony last week, Browne received the garden's highest award, the Henry Shaw medal, for transforming the "corporate culture" of British Petroleum and aligning it "with the goals of sustainability".

Peter Raven, the garden's director, described Browne as "perhaps the only fossil-fuel executive to speak decisively and meaningfully for corporate responsibility". In a speech at Stanford University last May that was widely seen as distancing the company from other leading members of the oil industry, Browne characterized the possibility of human influence on the climate as too significant to be ignored.

Pellat takes top job in French atomic energy

[PARIS] René Pellat, who is 63, last week succeeded Robert Dautray as high commissioner of the French Atomic Energy Commission (CEA), becoming the most

senior adviser to the government on nuclear affairs. Pellat, a research director at the Centre National de la Recherche Scientifique (CNRS), was president of the administrative council of CNRS between 1989 and 1992, and president of the French space agency, CNES, from 1992 to 1995. He has also been a member of the government's scientific council on defence matters, and of the scientific board of France Télécom. Pellat's research expertise is in plasmas, planet formation, and magnetic and inertial nuclear fusion.

Sanctuary charged with ill-treating chimps...

[WASHINGTON] The US Department of Agriculture (USDA) has filed charges against the Coulston Foundation, an organization based in New Mexico that cares for about half of the research chimpanzees in the United States (see *Nature* 390, 321; 1997), alleging breaches of the Animal Welfare Act. Similar charges were brought against Coulston in 1995. Since then two chimpanzees have died in controversial circumstances, and animal rights activists have waged a war of words against the New Mexico centre. "We're gratified that USDA has acted on the overwhelming evidence of negligence involved in the death of the chimps," says Suzanne Roy, director of In Defense of Animals.

...and Kyoto's chimp 'linguist' is pregnant



[TOKYO] Kyoto University's Primate Research Institute announced last week that Ai, its 21-year-old language-trained chimpanzee (above) has become pregnant through artificial insemination and is expected to give birth in August. The institute began trying to artificially inseminate Ai, which can recognize numerous Kanji (Chinese characters), numbers and visual symbols, in 1995, as part of a project to discern whether the use of symbols and tools can be transmitted across generations.

Tetsuro Matsuzawa, a professor in comparative cognitive science and the leader of the project, hopes that the birth of the baby chimpanzee will make it possible to determine whether Ai and her offspring can communicate by means of artificial language.

Scapegoat for fraud in Germany?

Sir — You have published articles alluding to scientific fraud in Germany (*Nature* **387**, 750 & **389**, 105; 1997). Because my name was mentioned prominently on each occasion, your readers should be aware of some additional facts.

Last year I was accused in private of having been involved in falsification of scientific data. I immediately wrote a full explanation, tantamount to a confession, and then offered to resign my position at the University of Lübeck. The resignation was accepted in June 1997. I do not believe that further victimization is appropriate.

Commissions of investigation have been established by academic authorities in Berlin, Ulm, Freiburg and Lübeck. In each case the commission sat in camera, sought information from me and others on an informal basis and encouraged no formal legal representation for the accused. The commissions issued selective press releases

but the full findings have not been sent to me or been made public.

I am forced to the conclusion that each commission met with the intention of limiting damage to the German academic community rather than discovering the full extent of culpability. It was only too easy to attack the person who had confessed while ignoring the evidence of greater wrongdoing that would require more rigorous investigation.

The resignation from my academic position at the University of Lübeck constituted part of a legal agreement by which I was offered a severance payment equivalent to almost one year's salary. The contract was signed by a representative of the Ministry of Science, Research, Education and Culture of the state of Schleswig-Holstein. In the event, the state government has reneged on its side of the agreement, claiming that its signatory

was not authorized to sign and that the agreement is therefore void. I, however, am expected to honour my part of the bargain.

In summary, I have very readily admitted to having been involved in falsification of scientific papers, an achievement of which I am not proud. I resent, however, the fact that, because I alone have admitted my mistakes at an early stage of the investigations, official bodies have found it expedient to imply that I was the major or conceivably the only culprit and that I was responsible for false data appearing in numerous papers, on some of which I was not even a co-author.

Marion A. Brach

*c/- Walter F Kalthoff,
Pacellistr. 14,
D-80333 München,
Germany*

AIDS therapy in Brazil

Sir — In your article "Better adherence vital in AIDS therapy" (*Nature* **390**, 326; 1997) you stressed the importance of persuading HIV patients to follow the complex new drug treatments.

We agree, but our data show that other factors may also be of importance, as shown by experience in Brazil where, since 1996, the government has made combined therapy available to all Brazilians with AIDS.

The Ambulatorio da Providencia Outpatient Clinic and Support House in Rio de Janeiro, run by the Roman Catholic Church, has worked on HIV infection since 1985. During that time, there has been a constant increase of HIV infection among our patients. The clinical population profile encompasses the poor and the 'social outcasts' of the city: slum (*favela*) dwellers, street children, sex workers, transsexuals and beggars. Some 700 HIV-positive/AIDS patients are followed up by the clinic and most of them are homeless or living in *favelas*. Some 25 homeless AIDS patients can be housed at the Santo Antonio Support House.

In 1990 we started a prospective study comparing survival time between homeless and formerly homeless housed individuals with AIDS in Rio de Janeiro. HIV-1-positive homeless subjects, HIV-positive volunteers who previously lived on the streets and agreed to live at our support house, and a control group of people living with relatives and who came from the poorest section of Rio society were followed up to 1997.

Survival time was calculated from the

date of AIDS diagnosis until the date of death. We divided the study into two phases. The first phase was from 1990 to 1995, when combined therapy was not available (we used monotherapy and/or double therapy) and the second from 1996 when triple therapy was made available by the government.

In the first phase, the mean survival of the homeless group (27 patients) was 8.2 months (range 1 to 33), in the formerly homeless group (26 patients) 17.8 months (1 to 48) and in the control group (59 patients) 18.3 months (2 to 60). The results from the log-rank test for pairs of survival curves revealed significant statistical differences between the groups of homeless and formerly homeless subjects ($P = 0.0018$) and the groups of homeless and housed subjects ($P = 0.0001$). In contrast, there was no significant statistical difference between the survival curves for the groups of formerly homeless and housed subjects ($P = 0.9704$). Better nutrition and hygiene, more frequent medical and psychological care together with controlled medicine intake, occupational therapy, decrease in promiscuity, alcohol and drug abuse and the provision of religious support may explain the difference in survival time observed in this phase of the study.

The second phase of the study has shown no deaths since the beginning of triple therapy (mean 13 months) for the formerly homeless (10) and housed (12) patients; all the patients are clinically well, CD4 counts have increased and viral loads are not detectable or below 700 copies (indicating efficacy of therapy). In contrast,

the survival time of homeless subjects has remained the same as that observed previously, because we do not give combined therapy if we cannot monitor adherence to treatment.

The use of combined therapy has increased the cost of each patient on therapy at the support house. A new problem has emerged because the number of clinically well patients discharged from the house has been very low, consequently increasing the list of HIV-positive homeless waiting to be taken into the support house.

These findings highlight the severe impact of the HIV-1 epidemic and its treatment in people from the poorest part of our society. The results of this study should be the reason for some cautious optimism by governments and encourage them to pursue measures to deal with longer life of patients and resocialization of social outcasts.

Luiz R. R. Castello-Branco

*Ambulatorio da Providencia,
Rua Francisco Eugenio 348,
Sao Cristovao,
Rio de Janeiro 22.941-120,
and Depto de Imunologia - FIOCRUZ,
Av. Brasil 4365,
Rio de Janeiro 21.045-900, Brazil
e-mail: branco@gene.dbbm.fiocruz.br*

Maria I. Linhares-Carvalho

*Ambulatorio da Providencia
e-mail: inez@pontocom.com.br*
Antonio C. M. Ponce de Leon
*Depto de Estatística, UERJ,
R Sao Francisco Xavier 524,
Rio de Janeiro, Brazil
e-mail: deleon@vmesa.uerj.br*

Arguments in favour of the space station

Sir — The Microgravity Advisory Committee (MAC) of the European Space Agency has read with interest your leading article and Briefing about the International Space Station (ISS)¹.

The MAC appreciates your efforts to clarify aspects of the scientific research planned on board the ISS and the technical and 'political' motivation for such an important international programme, but both the article and Briefing seem to show a negative attitude towards the project.

In both the physical and life sciences, research on phenomena and processes at near-zero gravity levels cannot be done on Earth for sufficiently long periods. Moreover, the necessity to repeat experiments at a pace consistent with that of the on-ground research — considered a basic requirement — can be assured only on board space stations.

The conclusion of your leading article that "the [scientific] community has placed its negative scientific judgements of ISS on the record" sounds unfair because it is not well supported by evidence: although it is true for some scientific societies, it does not represent the general opinion of the scientific community as a whole.

First, of the many personal opinions given by scientists interviewed by *Nature*, few positive comments seem to have found their way into print.

Second, in the ELGRA Report on microgravity research in Europe (September 1995), a committee consisting of a Nobel prizewinner and leading scientists recognized that "low-gravity is a useful tool — in some cases a unique tool — for the study of a number of physical and physico-chemical phenomena which are important in science, engineering and technology.... If the space station becomes available, the scientific community will make the best possible use of it to perform experiments." But, when writing of a "broad consensus" about the expected weak impact of the research planned for the ISS, such a statement is valid only if objective methods are used to establish it.

The experiments are being, and will be, chosen on the basis of rigorous peer-review procedures clearly stated in each announcement of opportunity. The only criterion has been and will be scientific excellence, a point on which there does exist a broad international consensus. Cost is not among the points the peers are asked to evaluate, for good reason. As correctly stated, the costs of these experiments, after deducting the costs of the orbiting laboratory, are not basically different from

the corresponding experiments performed on Earth.

Research on gravity-related phenomena in life and physical sciences has, in the past 20 years, given important results despite the scarcity of flight opportunities: these include advances in the understanding of the stability and dynamics of liquid interfaces (liquid columns, Marangoni motions), of transport phenomena (diffusion, Soret effects, crystal growth), of many physiological processes (the treatises on respiratory physiology must be considerably revised in the wake of many flight results) and in cell biology.

The possibility of gaining exciting insights from the 'continuous' use of an orbiting laboratory represents a real 'quantum jump' in scientific development and a potent magnet to attract young researchers and top-level scientists, as has already happened in the past few years.

The ISS represents one of the most demanding and challenging efforts made in the field of international scientific and technical cooperation. Whatever the other reasons for this programme, the situation is much the same as for any project of similar size: it has an impact on every aspect of science and technology precisely because of its size and the intensive and multidisciplinary collaboration it requires.

No scientific endeavour can, from the start, predict what benefit it may eventually have: let great projects grow and answer the question posed by Benjamin Franklin: "Of what use is the baby you have in your arms, Madam?"

Alberto Passerone
(MAC Chairman)

ICFAM-CNR,
Via de Marini 6, I-16149 Genova, Italy
e-mail: passerone@icfam.ge.cnr.it

Sir — Your discussion of the International Space Station¹ failed to draw attention to two of the most important arguments in favour of this ambitious project.

One is the development of new institutional arrangements needed for the management of complex international space projects. The quotation from John Logsdon (ref. 1, p. 734) that "in effect, an international space agency has been created for the station" may be overstating things at present, but there are strong reasons for believing that, if humanity is to have a significant future in space, a World Space Agency of some sort will be both necessary and desirable². If experience with building and operating the ISS helps to develop the institutional foundations for a future world space programme, that alone will be one of its most important legacies.

The other main argument in favour of the ISS is the experience that it will provide in building large structures in space. As your

discussion highlighted, many scientists have grave doubts about the suggested scientific applications of the ISS itself. But considerable, if long-term, scientific advantages are likely to follow from an ability to construct large structures in Earth orbit and beyond (for example, large space telescopes, lunar and planetary outposts and, eventually, interstellar space probes^{3,4}).

These arguments make sense, of course, only if one accepts that a significant human presence in space is a desirable future goal. Many in the scientific community do not accept this assumption, and the editorial pages of *Nature* have reflected this widespread scepticism. However, accepting that many of the arguments for an ambitious human space programme are social and political (reviewed in ref. 2) rather than narrowly scientific, it seems clear that science would be a major beneficiary. In which alternative future would we be likely to learn the most about the Universe and our place within it? And if we aspire to follow the latter course, it is not too soon to start laying the foundations for the technical and institutional infrastructure that will be required.

I. A. Crawford

Department Physics & Astronomy,
University College London,
Gower Street, London WC1E 6BT, UK
e-mail: iac@star.ucl.ac.uk

Sir — International Space Station "Worry no. 4" (manpower) is not a worry! I consider it a positive sign that "many researchers admit they worry that the station will not live up to its potential". This is a measure of their eagerness to use the station to conduct unique research in the microgravity environment. One of your worries is, however, misstated in your Briefing¹.

It implies incorrectly that insufficient crew time is available for research once construction is complete. It is true, and understandable, that crew time for research is limited during construction. But, when assembly is complete, the available crew time for research jumps to 160 hours per week, equivalent to four full-time crew members. Our best analyses show that this amount is not the constraining resource for the planned research.

The preservation and improvement of the research capability of the space station is a high priority of this programme.

John-David F. Bartoe

(Research Manager,
International Space Station)
NASA Johnson Space Center,
Houston, Texas 77058, USA
e-mail: john-david.f.bartoe1@jsc.nasa.gov

1. *Nature* **391**, 721 & 732–737 (1998).

2. Crawford, I. A. *Space Policy* **11**, 219–225 (1995).

3. Crawford, I. A. *Q. J. R. Astr. Soc.* **31**, 377–400 (1990).

4. Mallove, E. & Matloff, G. *The Starflight Handbook* (Wiley, New York, 1989).

A sense of direction

David Ferster

In order to sense visual motion, the brain must determine not only where an object is, but where it was a few moments ago. To do so, the brain may depend in part on a small number of curiously shaped neurons deep in the visual cortex.

For over 100 years, neuroanatomists have suffered from a minor obsession with a group of unusual neurons that lie in the bottom layers of the visual cortex. The Meynert cells (cell 'a' in Fig. 1) are large pyramidal neurons, distinguished by a set of long basal dendrites that project laterally in one direction for as much as 0.7 mm (ref. 1). Because they stand out so strikingly from standard-issue pyramidal cells, which come with shorter, more symmetrical dendrites, the Meynert cells' unusual morphology has long been the subject of speculation^{2,3}. Now, in a paper just published in *Neuron*⁴, Margaret Livingstone proposes by far the most compelling and specific function for the Meynert cells to date: that their asymmetrical dendrites make Meynert cells sensitive to visual motion, and in particular to motion in one direction.

To detect motion, the cortex must determine not only the position of a visual stimulus, but also its history. Direction selectivity, then, is a specific and experimentally approachable instance of how the cortex deals with time. In one of the simplest models of direction-selective neurons⁵, a cell receives two sets of synaptic inputs, one excitatory and one inhibitory, representing two adjacent regions of the visual field (Fig. 2, left; overleaf). The inhibitory input, however, has a longer latency (say, 25 ms longer) than the excitatory input. That is, the inhibitory post-synaptic potential (IPSP) evoked by a stimulus flashed in the right half of the receptive field arrives at the cell later than does the excitatory postsynaptic potential (EPSP) evoked by a stimulus flashed to the left.

Now consider a bar approaching the receptive field from the left (Fig. 2, upper right). The bar first triggers the EPSP, which in turn evokes action potentials in the cell. Later, the bar moves on to the inhibitory region of the receptive field, but the resulting IPSP is too late to prevent the EPSP from evoking action potentials. A bar approaching from the right, however, has a very different effect (Fig. 2, lower right). If the velocity is in the right range, the bar will encounter the right side of the receptive field about 25 ms before it reaches the left side. The time it takes the stimulus to cross the receptive field allows the IPSP time to reach its peak just as the EPSP does. The IPSP can therefore prevent the EPSP from generating action potentials, and the cell fails to respond to leftward motion of the stimulus.

Livingstone⁴ found just such an arrangement in direction-selective neurons recorded from the visual cortex of awake macaque monkeys. A bar flashed in one part of the receptive field evoked action potentials. In the adjacent region, the bar inhibited the cell's activity, but only after a delay. And, in accordance with the model, the velocity at which the cell was most direction selective was close to the spatial displacement of the two parts of the receptive field, divided

by the relative delays of the excitatory and inhibitory responses to the flashed bar⁶.

That the inhibition appears to be delayed relative to the excitation is not surprising: IPSPs in cortical cells are often slow to develop and can last over 100 ms. But what about the spatial offset between the excitatory and inhibitory inputs?

Enter the Meynert cells. Livingstone noticed that some of her direction-selective cells were located deep in the cortex where Meynert cells also lie. That, and previous hints that Meynert cells might be direction selective^{7,8}, led Livingstone to suspect that some of her cells were in fact Meynert cells, which in turn led her to speculate about the function of the Meynert cells' distinctive shape. Key to her theory is the cortical retinotopic map: each point in the cortex receives input from a corresponding point in the visual field such that the two-dimensional cortical surface forms an orderly map of the

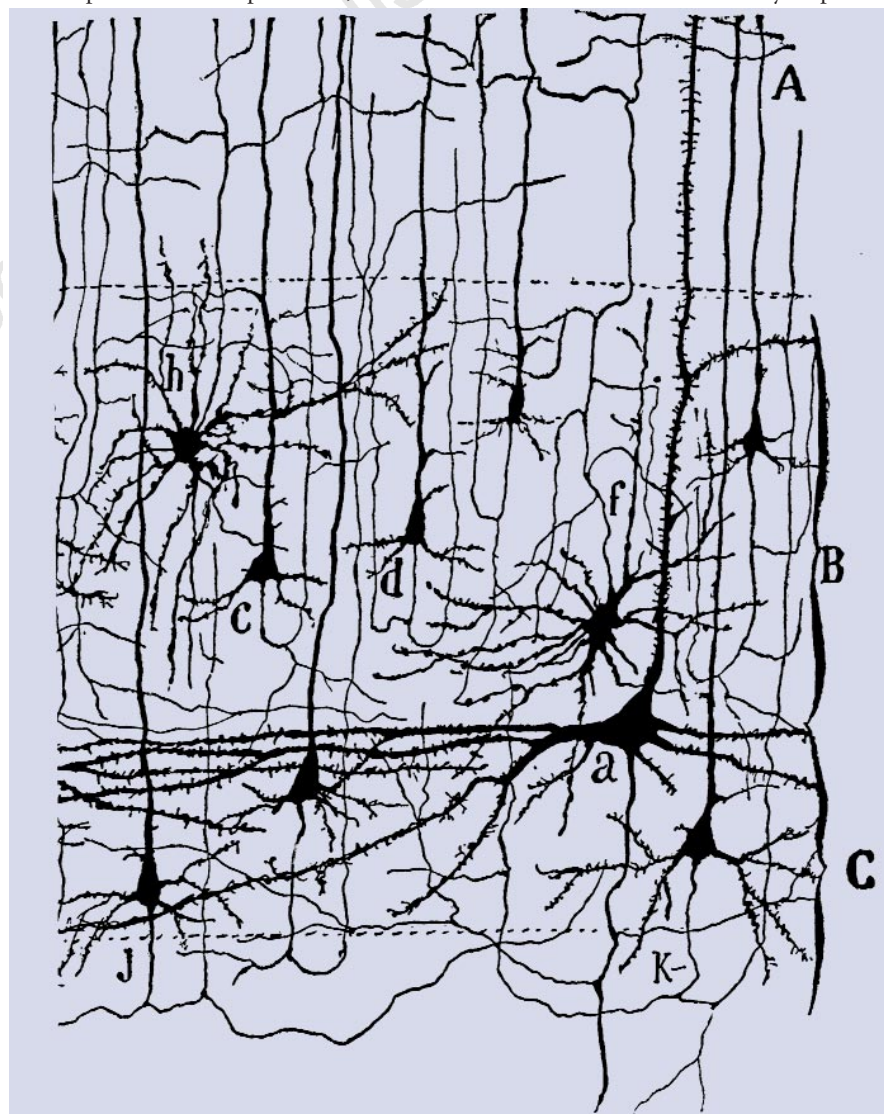


Figure 1 A drawing by Cajal¹ of neurons from the human cerebral cortex. Ordinary pyramidal cells (c, d, j, k) are distinguished by their triangular cell bodies, single upward-projecting apical dendrite and relatively symmetrical spray of laterally projecting basal dendrites. By comparison, the Meynert cell (a) has much longer and asymmetrically projecting basal dendrites, which Livingstone⁴ suggests contribute to the Meynert cells' direction selectivity.

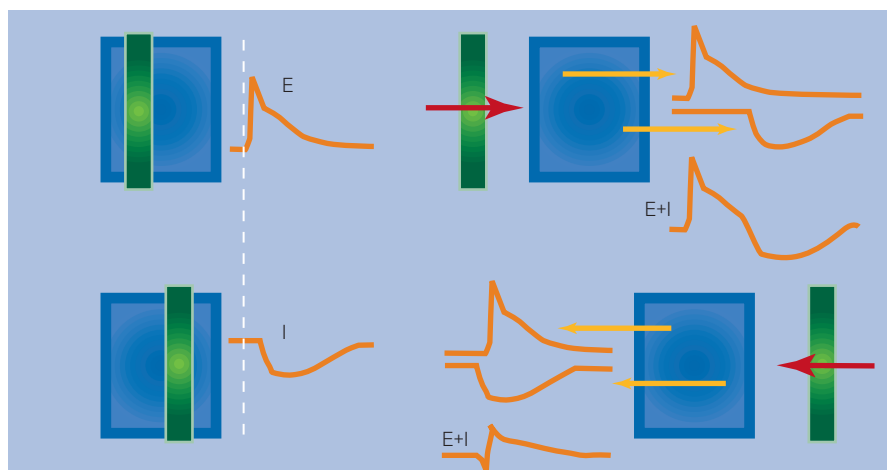


Figure 2 The possible organization of a direction-selective cell. (From ref. 5.) Flashing a bar in the left side of the receptive field (upper left) evokes an excitatory postsynaptic potential (EPSP), whereas flashing the same bar in the right side of the receptive field (lower left) evokes an inhibitory postsynaptic potential (IPSP), but with a longer delay. When a moving bar encounters the receptive field from the left (upper right), it first triggers the EPSP from the left side of the receptive field, and only later, the IPSP from the right side. The EPSP and IPSP sum at the cell body (E+I), but the IPSP is too late to prevent the EPSP from evoking action potentials in the cell. The bar approaching from the right, however, has a very different effect (lower right). If the bar's speed is in the right range, it will encounter the right side of the receptive field about 25 ms before it reaches the left side, giving the IPSP time to reach the cell in synchrony with the EPSP. The resulting sum of the EPSP and IPSP (E+I) is much smaller than for leftward motion, so that the IPSP effectively prevents the EPSP from generating action potentials in the cell.

two-dimensional visual field. Because the dendrites of Meynert cells project for such a long distance within the map, their tips, where many excitatory inputs are located, could sample synaptic inputs representing a different part of the visual field than does the cell body, where inhibitory inputs are preferentially located.

To test this idea, Livingstone determined the position of a direction-selective cell within the retinotopic map by recording the receptive field locations of nearby non-directional cells. In line with her proposal, the receptive fields of the nearby cells were nearly coincident with the inhibitory region of the directional cell's receptive field. The excitatory portion of the receptive field was therefore displaced in the map, presumably in the direction of the long basal dendrites.

Not all of the evidence fits completely with Livingstone's scheme. Winfield *et al.*⁹ found that the dendrites of neighbouring Meynert cells tend to line up in the retinotopic map. Yet Livingstone's cells don't all prefer the same direction of motion. Winfield *et al.* also found that the dendrites project parallel to ocular dominance columns and therefore most likely perpendicular to orientation columns, which raises a question about how the cells maintain their orientation selectivity. But these observations will no doubt be re-examined in light of Livingstone's model. Finally, not all of Livingstone's cells were located in the deep layers of the cortex. But other layers might contain neurons with asymmetrical dendrites, reminiscent of the Meynert cells.

As well as the temporal delay between the excitation and inhibition, Livingstone finds a gradual shift in response latency across the excitatory region of the receptive field¹⁰, the longest delays occurring in the regions farthest from the inhibitory zone. Interactions among the EPSPs could then contribute to direction selectivity in much the same way that the interactions between the EPSPs and IPSPs do: in response to a stimulus of the preferred direction, the EPSPs from different parts of the excitatory zone would reach the cell simultaneously and sum; with a non-preferred stimulus the EPSPs would be dispersed in time, making them less effective at

firing the cell. Livingstone attributes the spatial gradient in EPSP latencies to the time it would take the synaptic potentials to propagate from the tips of the long dendrites to the cell body. But whether the propagation speed of the dendrites is slow enough to account for the relatively long synaptic delays remains to be tested.

Distinct, spatially and temporally offset synaptic inputs reminiscent of those found by Livingstone have long been observed in studies of direction-selective neurons in cat visual cortex^{6,10–14}. If Livingstone's cells really are Meynert cells, or are neurons in other layers with asymmetrical dendritic projections¹⁵, her model could deepen our understanding of the mechanisms underlying direction selectivity. And Meynert cells would become one of the few neurons in the cortex whose distinctive dendritic morphology could be assigned a specific visual function. □

David Ferster is at the Institute for Neuroscience, Northwestern University, Evanston, Illinois 60208, USA.

e-mail: ferster@nwu.edu

1. Cajal, S. R. *Rev. Trim. Micrographica* **4**, 1–63 (1899).
2. Lund, J. S. *J. Comp. Neurol.* **147**, 455–496 (1973).
3. Winfield, D. A. & Powell, T. P. S. *Brain Res.* **277**, 223–229 (1983).
4. Livingstone, M. S. *Neuron* **20**, 509–526 (1998).
5. Barlow, H. B. & Levick, W. R. *J. Physiol. (Lond.)* **78**, 477–504 (1965).
6. McLean, J. & Palmer, L. A. *Vision Res.* **29**, 675–680 (1989).
7. Lund, J. S., Lund, R. D., Hendrickson, A. E., Bunt, A. H. & Fuchs, A. F. *J. Comp. Neurol.* **164**, 287–303 (1976).
8. Movshon, J. A. & Newsome, W. T. *J. Neurosci.* **16**, 7733–7741 (1996).
9. Winfield, D. A., Neal, J. W. & Powell, T. P. S. *Proc. R. Soc. Lond. B* **217**, 129–139 (1983).
10. Reid, R. C., Soodak, R. E. & Shapley, R. M. *Proc. Natl Acad. Sci. USA* **84**, 8740–8744 (1987).
11. Emerson, J. A. & Bergen, J. R. & Adelson, E. H. *Vision Res.* **32**, 203–218 (1992).
12. Saul, A. B. & Humphrey, A. L. *Vis. Neurosci.* **8**, 365–372 (1992).
13. Albrecht, D. G. & Geisler, W. S. *Vis. Neurosci.* **7**, 531–546 (1991).
14. DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. *J. Neurophysiol.* **69**, 1118–1135 (1993).
15. Elston, G. N. & Rosa, M. G. *Cereb. Cortex* **7**, 432–452 (1997).

Bose–Einstein condensates

Superfluids mixing it up

Brett D. Esry and Chris H. Greene

Imagine two fierce competitors tossed into a trap, locked in a struggle with each other for supremacy. The quantum-mechanical match is hardly fair: the first competitor, 50,000 hefty ⁸⁷Rb atoms in a Bose–Einstein condensate (BEC), outweighs the second almost fourfold. The svelte opponent of the rubidium BEC is an equal number of lighter, smaller ²³Na atoms that are also Bose–Einstein condensed. The entire cloud hovers magnetically, a whisper above absolute zero on the temperature scale. New theoretical work by Pu and Bigelow¹ predicts striking properties for such a two-component BEC formed from sodium and

rubidium atoms.

In theory^{1–4}, a regime might exist in which superfluid condensates interpenetrate like miscible fluids. Depending on the strength of interactions between two unlike atoms, however, a very different phase-separated regime could arise in which the sodium atoms form a shell that encloses the rubidium. In the latter case, a higher-energy metastable state has now been predicted¹, in which the rubidium atoms form a shell around the periphery of an inner sodium cloud. Pu and Bigelow also point out an intriguing experimental possibility: the inner cloud can be intentionally

squeezed into a smaller volume, simply by adding more atoms to the outer shell (Fig. 1).

The fun with two-component Bose–Einstein condensates began in 1997, when Myatt *et al.*⁵ observed a double condensate composed of two clouds of ⁸⁷Rb atoms in different internal spin states. The experiment astonished the BEC community, because it violated an expectation gleaned from a quarter of a century of research into spin-exchange processes: atoms colliding in these states should scatter into different spin states, in which they would be ejected from the trap by magnetic forces. But fortunately, the dastardly spin-exchange process happens to be suppressed in ⁸⁷Rb by nearly two orders of magnitude.

Only this accident permits the Myatt *et al.* double condensate to form and to persist^{6–8}. So as long as multiple-component condensates are trapped magnetically, they will have to be ⁸⁷Rb — unless another ‘accident’ occurs for a different alkali atom. But Ketterle and colleagues have demonstrated BEC in an optical trap⁹, in which focused light captures the electric dipole moment of each atom rather than its magnetic moment. A trap of this type will hold an atom even after a spin-changing collision; potentially, a richer variety of multiple-component condensates can therefore be studied.

At these ultra-low temperatures, the scattering length A controls the strength of

interactions between any two particles. (The Newtonian-minded reader can view A very roughly as the sum of the radii of two colliding billiard ball atoms.) In any bargain-base condensate with a single component, every atom is identical and there is only one scattering length. For a double condensate, however, three scattering lengths are needed: A_{11} (Rb–Rb, for the system studied by Pu and Bigelow), A_{22} (Na–Na) and A_{12} (Rb–Na). The first two are accurately known, but unfortunately A_{12} (Rb–Na) has never been measured. All we know is that it lies somewhere between plus and minus infinity. (Less uncertainty exists about the properties predicted¹⁰ for a double condensate composed of the two isotopes ⁸⁵Rb and ⁸⁷Rb.)

To cope with this vast uncertainty, Pu and Bigelow solved the problem for many different values of A_{12} . Three phases could occur^{1–4}: (i) a miscible, fully interpenetrating phase, if A_{12} (Rb–Na) is between -3.4 nm and $+3.4$ nm; (ii) an immiscible, phase-separated double condensate if $A_{12} > 3.4$ nm; (iii) no condensate at all, if $A_{12} < -3.4$ nm. In the last case the interaction is too attractive and the double condensate undergoes a runaway collapse, as it tries to form clusters of sodium and rubidium, or ‘metallic snowflakes’.

Researchers have longed for years to study interpenetrating superfluids, but the only superfluids available — ³He and ⁴He — steadfastly refuse to mix when cooled togeth-

er. Should the sodium–rubidium interaction A_{12} turn out to be in the right range, this will be the first example of miscible superfluids. Yet interesting science beckons even if $A_{12} > 3.4$ nm, because two curious effects can still be explored in this phase.

The first follows from the tendency of the sodium atom cloud to form a shell around the outside of the rubidium atom condensate cloud. Pu and Bigelow¹ suggest that by increasing the sodium cloud size, a higher pressure could be brought to bear on the inner rubidium cloud, causing it to shrink, increasing inelastic collision rates enormously. The second effect possible in this phase follows from the smallness of the energy cost of swapping each sodium atom with a rubidium atom. The swap results in a metastable state with rubidium atoms packed around an inner sodium cloud; it has slightly higher energy than the ground state, and a long lifetime against decay into the true ground state.

This regime may have another peculiar property. The condensate is predicted to be unstable¹, an instability that has been shown to lead to a spontaneous symmetry breaking in other cases^{11,12}. Our calculations confirm this: the immiscible phase should not possess the spherical symmetry of the trap, but should instead consist of two unmixed superfluids arranged more or less linearly along some axis in space¹².

But the notion that multiple-component BEC experiments are constrained by what nature has handed us, in the way of atom–atom scattering lengths, appears to be far too pessimistic. Just last month, for instance, Wolfgang Ketterle’s group reported the ability to tune a sodium–sodium scattering length by varying the magnetic field across a Feshbach resonance¹³. We seem to be at the threshold of a new era of designer multi-condensates. □

Brett D. Esry is at the Institute for Theoretical Atomic and Molecular Physics, Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA.
e-mail: besry@cfa.harvard.edu

Chris H. Greene is in the Department of Physics and JILA, University of Colorado, Boulder, Colorado 80309-0440, USA.

e-mail: chg@jilacg.colorado.edu

1. Pu, H. & Bigelow, N. *Phys. Rev. Lett.* **80**, 1130–1133; 1134–1137 (1998).
2. Law, C. K. *et al.* *Phys. Rev. Lett.* **79**, 3105–3108 (1997).
3. Ho, T.-L. & Shenoy, V. B. *Phys. Rev. Lett.* **77**, 3276–3279 (1996).
4. Esry, B. D. *et al.* *Phys. Rev. Lett.* **78**, 3594–3597 (1997).
5. Myatt, C. J. *et al.* *Phys. Rev. Lett.* **78**, 586–589 (1997).
6. Julienne, P. S. *et al.* *Phys. Rev. Lett.* **78**, 1880–1883 (1997).
7. Kokkelmans, S. J. J. M. F. *et al.* *Phys. Rev. A* **55**, R1589–R1592 (1997).
8. Burke, J. P. Jr *et al.* *Phys. Rev. A* **55**, R2511–R2514 (1997).
9. Stamper-Kurn, D. M. *et al.* *Phys. Rev. Lett.* **80**, 2027–2030 (1998).
10. Burke, J. P. Jr *et al.* *Phys. Rev. Lett.* **80**, 2097–2100 (1998).
11. Ohberg, P. & Stenholm, S. *Phys. Rev. A* **57**, 1272–1279 (1998).
12. Esry, B. D. Thesis, Univ. Colorado (1997).
13. Inouye, S. *et al.* *Nature* **392**, 151–154 (1998).

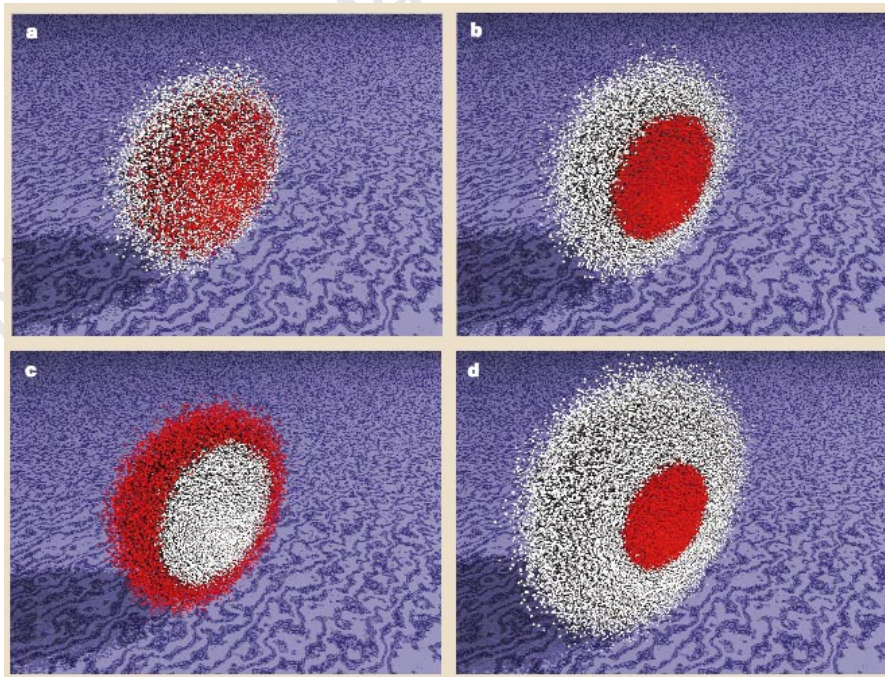


Figure 1 Composite condensates: the distribution of a spherical two-component Bose–Einstein condensate made of ⁸⁷Rb atoms (red) and ²³Na atoms (white), at absolute zero. Half of the condensate has been cut away to expose the centre. a, Miscible superfluid of 50,000 rubidium and 50,000 sodium atoms, with a rubidium–sodium scattering length $A_{12} = 1.8$ nm. b, Immiscible superfluid, with rubidium atoms inside a shell of sodium atoms; $A_{12} = 9.6$ nm. c, As b, but the higher-energy metastable state is shown, with the sodium on the inside. d, As b, but now 200,000 sodium atoms compress 50,000 of rubidium into a far smaller volume. These pictures are from our simulations, with parameters the same as those used by Pu and Bigelow¹.

NATURE

100 YEARS AGO

The *British Medical Journal* for March 19 contains an important paper by Dr. Luigi Sambon, on the "Etiology of Sunstroke." Dr. Sambon adopts what at first appears a somewhat startling theory, namely, that sunstroke is not due to excessive heat or exposure to the sun, but is an infectious disease due to a specific organism. The author's case rests on three lines of argument. He begins by showing that excessive heat does not produce the disease; stokers, miners, and iron-workers are exposed to temperatures higher than those of any known climate. ... Dr. Sambon next discusses the geographical distribution of the disease, and proves that the areas in which it is endemic are strictly defined. It is very common in the low-lying regions of the Eastern United States, between the Appalachians and the Atlantic; it is unknown in Europe; it extends along the Nile Valley, Red Sea, and Persian Gulf ... Another peculiar feature of the disease explicable on the infection theory is the occurrence of epidemics, which may decimate hospital wards and not affect men exposed to greater heat and sun. Dr. Sambon concludes that the distribution, etiology, morbid anatomy, and epidemic character of the disease together demonstrate its organic origin. From *Nature* 31 March 1898.

50 YEARS AGO

... one of the few serious complaints raised to-day against the young scientist graduate refers to his alleged inability to express findings in a written report. ... The main trouble seems to arise from the fact that often the young scientific worker has no love for writing. Whatever disciplinarians may say, it is doubtful whether anyone makes a real success of a task which gives him no sort of satisfaction. Enthusiastic and hard-working experimenters show an obvious reluctance to produce reports by the appointed day, and require a surprising length of time to produce a very simple memorandum. Even if the results, when ultimately completed, were perfect, the obvious disinclination to use the pen in the communication of ideas reveals an unsatisfactory state of affairs. This aversion to writing is often shown by men who are extremely successful in explaining their ideas by word of mouth. From *Nature* 3 April 1948.

Molecular endocrinology

Steroids tickle cells inside and out

Didier Picard

On reaching a cell, an extracellular signal can either knock on the door (the plasma membrane), or it can just walk in. Most hydrophilic signals such as peptides are unable to cross the membrane, and they signal through specific, membrane-bound receptors. But steroid hormones are cholesterol derivatives and, being lipophilic, they are thought to cross membranes readily.

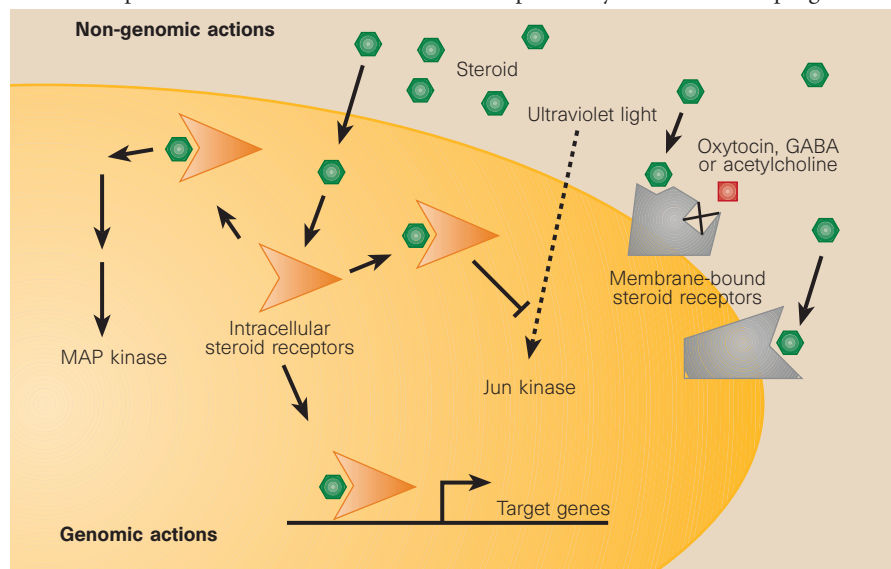
The first intracellular steroid receptors were discovered in the 1960s, and many receptors (along with other members of the intracellular receptor superfamily) were cloned in the 1980s and '90s. The central dogma of steroid signalling thus became that steroids signal through intracellular receptors, which are hormone-regulated transcription factors. But, in their excitement, researchers may have underestimated nature's wizardry — because steroids *can* reach specific receptors inside cells and elicit changes in gene expression ('genomic effects'), this does not mean that they always have to.

And they don't. On page 509 of this issue, Grazzini *et al.*¹ describe a transcription-independent signalling pathway of the steroid progesterone. Progesterone is essential for maintaining pregnancy in mammals, and it has the opposite effect to oxytocin, a nonapeptide that induces uterine contractions and may contribute to the onset of labour and parturition. Grazzini *et al.* show

that progesterone inhibits oxytocin signalling by binding to the membrane-bound oxytocin receptor. It binds with an affinity (20 nM) that is considerably lower than its binding to the intracellular progesterone receptor (<1 nM). But this is compatible with the astronomical concentration of progesterone during pregnancy (500 nM).

Does progesterone really act on the outside of the cell? Undoubtedly it does, because the authors found that progesterone tethered to a membrane-impenetrable carrier protein also worked. Progesterone binding reduced the number of oxytocin receptors that were available to bind oxytocin (and, thereby, relay the oxytocin signal into the cell). The oxytocin receptor belongs to the large class of membrane-bound receptors that relay their signals through guanine-nucleotide-binding (G) proteins to intracellular target proteins such as phospholipase C. Grazzini *et al.* found that progesterone inhibits two functional effects of oxytocin signalling: the production of the second messenger inositol-1,4,5-trisphosphate, and an increase in the concentration of intracellular Ca^{2+} . By recording the changes in the Ca^{2+} concentration, they showed that inhibition takes place in less than a minute and is readily reversible.

The pharmacology also held surprises and definitely put intracellular receptors off-side. The synthetic progesterone R5020 and the politically infamous anti-progesterone



either through membrane-bound receptors, they can modulate signalling of cin, as described by Grazzini *et al.*¹. Or they can elicit a p ntracellular receptors are ligand-regulated transcription factors, although non-genomic actions have also been reported. These include activation of the mitogen-activated protein (MAP) kinase signalling cascade⁵, and inhibition of the activation of Jun kinase⁶.

RU486 (the 'morning after' pill) also inhibit the rat oxytocin receptor, although other classes of steroids, including oestrogen and glucocorticoid, cannot. But the authors found that the human oxytocin receptor was not inhibited by progesterone or any of the other above-mentioned ligands. Instead, the inhibitory steroid for the human receptor turned out to be the progesterone metabolite 5 β -dihydroprogesterone.

How does it work? Progesterone could bind the oxytocin receptor at an allosteric effector site, inducing a conformational change that prevents oxytocin from binding to its own binding site. To test this, the molecular details need to be characterized and compared to those in other steroid-modulated systems. Indeed, this is not the first example of a non-genomic effect. As-yet uncharacterized receptors seem to mediate steroid signalling in a variety of biological processes², and both stimulatory and inhibitory effects of progesterone and its derivatives have been described on receptors for the neurotransmitters GABA, NMDA (N-methyl-D-aspartate) and acetylcholine²⁻⁴. The work of Grazzini *et al.*¹ extends this phenomenon from these ligand-gated ion channels to G-protein-coupled receptors. Moreover, the olfactory receptors in our nose belong to this class, so it is possible that sex-specific differences in the steroid milieu in the olfactory epithelium itself contribute to the proverbially keener sense of smell of women?

Steroids do not stop here. Once inside a cell, they affect signalling pathways in a transcription-independent fashion through their cognate intracellular steroid receptors. For example, the oestrogen receptor can activate the mitogen-activated protein kinase signalling pathway⁵ that is normally turned on by peptide factors. And the glucocorticoid receptor can interfere with the activation of a related signalling pathway that is involved in the response to ultraviolet and inflammatory signals⁶ (Fig. 1).

The bewildering diversity of steroid actions — both inside and outside cells — suggests that it should be possible to develop therapeutic drugs that specifically affect only one mode of action. For instance, compounds that mimic the inhibitory effect of progesterone (or 5 β -dihydroprogesterone in humans) could help to antagonize the effects of oxytocin in pre-term labour, but with fewer side-effects than the currently used beta-mimetic drugs. □

Didier Picard is in the Département de Biologie Cellulaire, Université de Genève, Sciences 3, 1211 Genève 4, Switzerland.

e-mail: Picard@cellbio.unige.ch

1. Grazzini, E., Guillon, G., Mouillac, B. & Zingg, H. H. *Nature* **392**, 509–512 (1998).
2. Wehling, M. *Annu. Rev. Physiol.* **59**, 365–393 (1997).
3. Baulieu, E.-E. *Recent Prog. Horm. Res.* **52**, 1–32 (1997).
4. Valera, S., Ballivet, M. & Bertrand, D. *Proc. Natl Acad. Sci. USA* **89**, 9949–9953 (1992).
5. Migliaccio, A. *et al. EMBO J.* **15**, 1292–1300 (1996).
6. Caelles, C., González-Sancho, J. M. & Muñoz, A. *Genes Dev.* **11**, 3351–3364 (1997).

Condensed-matter physics

Stripes of a different stripe

A. J. Millis

In some materials, the electronic charge density organizes itself into stripes. Stripe physics has been of intense interest to condensed-matter physicists, as an example of a non-trivial ordering phenomenon, and because of its possible connection to high-temperature superconductivity. On page 473 of this issue¹, Mori *et al.* report a novel paired stripe order. Their work opens a new window onto stripe physics and onto the fundamental issue of the interplay between hybridization and interactions.

Hybridization, the overlap of electron wavefunctions centred on different sites, is a quantum-mechanical effect that allows electrons to hop from one atom to another, thus tending to spread the electronic density uniformly through the solid. In contrast, interactions of electrons with one another and with displacements of the atoms in the solid tend to promote non-uniform charge distributions. If hybridization is dominant, a conventional metal (such as aluminium) or insulator (such as diamond) results. But if interactions are dominant, charge order-

ing may occur: in a charge-ordered state the electrons arrange themselves in a pattern with periodicity differing from that of

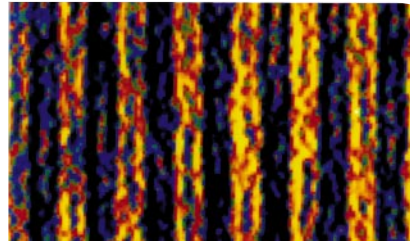
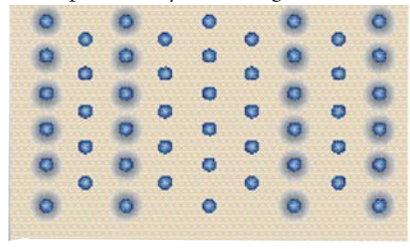


Figure 1 Stripe pairs in the manganite La_{1-x}Ca_xMnO₃. An electron diffraction image shows the overall structure of the stripe phase; the schematic inset shows a close-up. Itinerant electrons (diffuse clouds) may hop from manganese site to manganese site, but at low temperature are localized into pairs of lines.

the underlying lattice. The precise pattern is determined by the nature of the interactions and by the residual effects of hybridization.

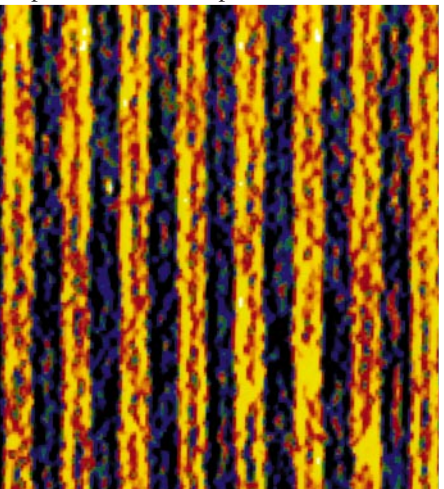
In many circumstances the charge ordering takes the form of stripes (Fig. 1), which seem to be the best compromise between the localizing effect of interactions and the delocalizing effect of hybridization. Along the stripe, the charge density is constant except for the variation imposed by the periodicity of the underlying lattice; perpendicular to the stripe, the charge density varies with a period different from that of the lattice.

Stripe order has been discovered in materials closely related to the high-temperature superconductors², and theorists have proposed that superconducting compounds should be regarded as 'quantum fluctuating stripe phases'³ in which the hybridization is not large enough to prevent stripes from forming, but does cause their positions to fluctuate so strongly that no static order occurs.

Stripe phases have been difficult to study quantitatively, in part because in most materials the balance between hybridization and interaction can be varied only by changing the chemical composition, which introduces a host of extraneous complications.

The new work of Mori and co-workers promises to change this situation. They studied La_{1-x}Ca_xMnO₃, one of the 'colossal magnetoresistance' manganites, so-called because an applied magnetic field can change their resistivity by a factor which can be as large as several thousand. That property might be technologically useful, but from the standpoint of basic physics the manganites' most interesting feature is that their quantum-mechanical hybridization can be varied over a wide range by applying pressure or a magnetic field.

The variation with pressure is simply explained: increased pressure moves atoms



S. MORI ET AL.

closer together, letting electrons hop more easily. The variation with magnetic field arises from a neat piece of atomic physics. An electron has an intrinsic angular momentum, or spin, whose axis may point in any direction in space. Each manganese atom in $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ has its own 'core spin', formed of three localized electrons with parallel spins, and may also carry 'itinerant electrons', which can hop from manganese to manganese. On a given site, an itinerant electron is constrained by the 'Hund's coupling' of atomic physics to have its spin axis parallel to that of the core spin. The consequence is that the amplitude for an electron to hop from one site to another depends on the relative orientation of the core spins at the two sites, being greatest when the core spins are parallel and least when they are antiparallel. Because an applied magnetic field tends to align the core spins, it increases the hybridization.

At low temperature, and in the composition range $x \geq 0.5$, $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ is a charge-ordered insulator. Mori *et al.* have discovered that this charge-ordering takes the form of stripes, and that the fundamental unit is surprisingly complex: it is a sort of sandwich, comprising a pair of stripes with one itinerant electron per manganese, bracketing a stripe of manganese sites with none (Fig. 1). Lattice distortions also occur, most notably a 'Jahn-Teller' displacement in which some Mn-O bonds become shorter than average and others longer. This basic three-row sandwich unit is repeated throughout the crystal, with the mean spacing between units set by the average charge density.

The discovery is important for several reasons. The structure of the stripes was not anticipated, so theorists will have to re-examine their models. Also, the charge density variation across the stripe is large, so properties of stripes may be more easily studied than in other compounds. Moreover, the manganite materials exist in a wide variety of crystal forms, so a detailed study of the effect of lattice structure and dimensionality on stripe formation should be possible.

But most importantly, the sensitivity of the hybridization to pressure and magnetic field should allow experimentalists to study its effects in detail. For example, we already know that the low-temperature insulating stripe phase may be converted into an electrically conducting phase by applying pressure or a magnetic field⁴. Presumably this metal-insulator transition is caused by increased quantum fluctuations coming from the increase of hybridization with pressure or field, and it will be interesting to see whether the resulting metal is a realization of the 'fluctuating stripe phase' postulated for high-temperature superconductivity. □

A. J. Millis is in the Department of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA.

e-mail: millis@pha.jhu.edu

1. Mori, S., Chen, C. H. & Cheong, S.-W. *Nature* **392**, 473–476 (1998).
2. Tranquada, J. M., Sternlieb, B. J., Axe, J. D., Nakamura, Y. & Uchida, S. *Nature* **375**, 561–563 (1995).
3. Emery, V. J. & Kivelson, S. A. *Physica C* **235–240**, 189–192 (1995).
4. Kuwahara, H. *et al. Phys. Rev. B* **56**, 9386–9396 (1998).

Biomathematics

Merging lines and emerging levels

Karl Sigmund and Eörs Szathmáry

Talk of the 'higher principles of life' has a ring of vitalism to it. So it must have taken Michael Polanyi considerable pluck to argue in 1968, at the onset of the spectacular advances of biochemistry and molecular genetics, that life has an 'irreducible structure'.

He wrote: 'We can recognize a strictly defined progression, rising from the inanimate level to ever higher additional principles of life. ... Evolution may be seen, then, as a progressive intensification of the higher principles of life'¹. This 'progress', nowadays described as a series of major transitions in evolution², is often due to the emergence of new units. We are used to seeing individuals as the units of selection, but they are groups of cells, which include, in turn, groups of organelles descended from some proto-cells, and whose nuclei are groups of genes.

Why should we then have qualms about viewing groups of individuals (societies,

colonies or species) as units on their own? The emergence of new levels of organization was the subject of a workshop at the École Normale Supérieure in Paris in January*.

The fact that this workshop was attended by evolutionary biologists and mathematicians in almost equal numbers reflects a significant trend. Biologists' favourite adopted science is chemistry, but chemistry does not tackle many basic problems, such as the co-evolution of hosts and parasites, the merger of lineages, or the tuning of mutation rates. It is *populations* of virus particles, or immune cells, or hosts, that regulate one another's frequencies. The feedback loops of these ecosystems are too complex to be understood by verbal arguments alone, and can only be analysed by mathematical means.

Both mathematicians and biologists have become more aware of the advantages of

talking to each other^{3,4}, but this association does not come naturally. R. A. Fisher, a founding father of population genetics, attributed the difficulties of understanding between mathematicians and biologists to their vastly different training⁵.

It can be taxing to listen to the other side, especially if you are 'fort en math' and accustomed to rank at the top of France's intellectual scale. On several occasions, eminent mathematicians have developed ambitious biological speculations without the help of biologists, and altogether failed to convince them. And, indeed, someone accustomed to seeing mathematics as the queen of science, rather than its servant, will find it difficult to work with biologists.

The new emphasis of biomathematics is on explaining large evolutionary transitions in terms of nonlinear, local interactions between replicating units. Such interactions often lead to coherent structures whose existence is strikingly obvious in computer simulations, but which cries out for a mathematical underpinning; and this needs expertise on bifurcations of dynamical systems, hydrodynamical limits of interacting particles, and nonlinear partial differential equations — areas in which France is currently leading.

One spectacular challenge for such experts is the hydrodynamical slime-mould model of F. Siegert (Univ. Munich), in which two-dimensional arrays of isolated single cells use chemical gradients to form a three-dimensional, multicellular aggregate with a twisted scroll wave running along its axis, which makes it move, comically, like a real slug⁶.

Other self-generated spatial correlations overturn conclusions derived under the assumption of completely mixed populations of hosts and parasites. In the completely mixed case, parasites tend to kill all the hosts and then die; but with spatial variation, self-replicating clusters of parasites and hosts can form (D. Rand, Univ. Warwick; M. van Baalen, Univ. Paris VI).

Spatial considerations also modify the concept of fitness, usually defined as individual reproductive success. In the case of the Prisoner's Dilemma, for instance, which is frequently used (possibly overused) as a toy model for the evolution of cooperation, fitness translates as the speed of travelling waves of defectors and tit-for-tat players, as these waves can overwhelm populations of the opposite type (R. Ferrière, ENS, Paris).

M. Boerlijst (Univ. Amsterdam) and P. Hogeweg (Univ. Utrecht) argued that the blobs and spirals seen in their computer simulations are emerging 'units of selection' protecting otherwise unstable feedback loops of self-replicating molecules, or of parasites and hosts. The blobs play the part of a protective membrane, as in a primitive multicellular organism. ►

* Units of Selection and the Major Transitions of Life, Paris, 30–31 January 1998.

Beyond the semantics raised by these hotly debated proposals lurks the problem of how to compare fitness differences across levels of selection, a problem that is unavoidable if one wants to understand the organization of a genome able to modify its own mutation rate (B. Godelle, Univ. Paris XI)⁷ or the transition from unicellular to multicellular eukaryotes — a transition that must cause smouldering conflicts within the developmental system of the emerging entity (R. Michod, Univ. Arizona).

A careful analysis of symbiosis and mutualism stresses that analysis at different levels — physiological, ecological and evolutionary — may lead to different conclusions (U. Dieckmann, Intl Inst. Appl. Syst. Anal.; R. Law, Univ. York). Even in the face of persistent physiological exploitation of one partner by the other, evolution can select for stable symbiotic structures: such adaptations lead to a kind of dependence that is more like addiction than mutual benefit.

This may shed light upon the origin of mitochondria in eukaryotic cells. Suggestions favouring a very early acquisition of mitochondria suffer from two unsolved problems: the method of acquisition (no sensible alternative to phagocytosis has ever been suggested), and the initial advantage of such an association. Perhaps proto-mitochondria were once parasitic, and only later evolved into ATP-generating slaves. As one

of us (E.S.) pointed out, isogametic sex may have been crucial, as it allows the spread of moderately harmful intracellular symbionts.

Scenarios of this type vastly expand the range of conditions under which separate lineages can be expected to merge into symbiotic units. This leads one to hope, on yet another level, that mathematicians and biologists will find their emerging association of mutual benefit. They may eventually become addicted to each other's company. □

Karl Sigmund is at the Institute for Mathematics, University of Vienna, Strudlhofgasse 4, A-1090 Vienna, Austria, and the International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria.

e-mail: ksigmund@esi.ac.at

Eörs Szathmáry is in the Department of Plant Taxonomy and Ecology, Eötvös University, Budapest, and the Collegium Budapest, Szentharomsag u. 2., H-1014 Budapest, Hungary.

e-mail: szathmary@colbud.hu

1. Polanyi, M. *Science* **160**, 1308–1312 (1968).
2. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Freeman, Oxford, 1995).
3. Fisher, R. A. *The Genetical Theory of Natural Selection* (Dover, New York, 1958).
4. Levin, S. A., Grenfell, B., Hastings, A. & Perelson, A. S. *Science* **275**, 334–343 (1997).
5. Fagerstroem, T., Jagers, P., Schuster, P. & Szathmáry, E. *Science* **274**, 2039–2040 (1996).
6. Dormann, D., Weijer, C. & Siebert, F. *J. Cell Sci.* **110**, 1831–1837 (1997).
7. Taddei, F. *et al. Nature* **387**, 700–702 (1997).

Conservation biology

Inbreeding leads to extinction

Richard Frankham and Katherine Ralls

Do genetic problems contribute to the endangerment and extinction of wild populations? Conservation biologists initially thought¹ that they would — and seriously so. But it is extremely difficult to demonstrate that inbreeding contributes to the extinction of wild populations. On page 491 of this issue, however, Saccheri and colleagues² provide the first direct evidence that it does, with their elegant work on a wild butterfly metapopulation in Finland.

Theoretical work in the 1980s indicated that small populations in the wild suffer from increased extinction because of an unavoidable increase in matings between close relatives. Inbreeding reduces reproductive success in populations of naturally outbreeding species, both in captivity^{3,4} and in the wild⁵, and it also increases extinction rates in laboratory populations of fruitflies and mice⁵. However, in an influential paper⁶, Lande argued that random demographic and environmental events will drive small wild populations to extinction before genetic factors come into play. Environmental events, ranging from annual variation in climatic variables (such as rainfall) to catastro-

phes (such as disease epidemics), do increase the probability of extinction. Furthermore, inbreeding typically interacts with demography by reducing fecundity, juvenile survival and lifespan. Because there is no direct evi-

dence that inbreeding contributes to extinction of wild populations, some researchers have continued to question the relevance of genetic factors^{7,8}.

The Glanville fritillary butterfly (*Melitaea cinxia*; Fig. 1) studied by Saccheri *et al.*² has a predictable yearly life cycle. Adults emerge, mate, and lay eggs in June. Caterpillars feed in conspicuous family groups of 50 to 250 individuals, then diapause (suspend development) from August until March of the following year, and resume feeding and pupate in May. The butterfly metapopulation consists of numerous small populations that breed in about 1,600 suitable dry meadows of different size and varying distance from one another. Some populations are very small, often consisting of the offspring of a single pair of butterflies. Consequently, populations in individual meadows often disappear, but many meadows are eventually recolonized, with an average of 200 extinctions and 114 colonizations per year.

Because small population size results in both inbreeding and loss of genetic variation, the degree of genetic variation in a population serves as a measure of the extent to which it is inbred. Saccheri *et al.* determined the genotypes of female butterflies from 42 populations at eight variable genetic loci (polymorphic loci). They sampled relatively large, non-isolated populations, as well as smaller, relatively isolated populations. The authors found that populations with less genetic variation were more likely to become extinct. Furthermore, multiple logistic regression showed that genetic diversity predicted extinction risk after accounting for all known demographic, ecological and environmental causes of extinction in this well-studied butterfly metapopulation. Inbreeding reduced the egg hatching rate and larval survival, lengthened the dura-



Figure 1 Doomed liaison — a mating pair of Glanville fritillary butterflies (*Melitaea cinxia*). From their studies of a metapopulation of this species, Saccheri *et al.*² found that inbreeding contributes to the extinction of wild populations.

I. SACCHERI

tion of the pupal period (so that inbred pupae were more likely to be parasitized), and shortened female lifespan (so that inbred females tended to lay fewer eggs). Overall, inbreeding explained 26% of the variation in extinction rate among the butterfly populations.

Several indirect lines of evidence imply that these results from the Glanville fritillary butterfly can be extended to other species. First, theoretical studies have shown that genetic factors probably contribute to extinctions, even when demographic and environmental fluctuations and catastrophes are operating⁹. Second, genetics may be a factor that makes island populations prone to extinction. Although humans (and the animals that they have introduced) have decimated many island populations, these island populations have lower genetic diversity than mainland populations¹⁰. Moreover, many are inbred to levels where captive populations show an increased risk of extinction from inbreeding¹¹. Third, ratios of effective-to-census population sizes seem to be much lower than suspected¹², so genetic concerns become more important in larger populations than previously believed. Fourth, endangered species tend to have lower genetic diversity than non-endangered species⁴ — this would not be expected if ecological factors drove populations to extinction

before genetic factors became important. Finally, the extinction rate of a wild plant was higher in experimental populations with low versus higher genetic variation when both were planted in the field¹³.

It is hard to escape the conclusion that genetic factors are involved in the extinction of wild populations. Consequently, genetic factors must be considered when assessing endangerment and devising recovery plans for threatened species. □

Richard Frankham is in the School of Biological Sciences, Macquarie University, New South Wales 2109, Australia.

e-mail: rfrankha@rna.bio.mq.edu.au

Katherine Ralls is in the Department of Zoological Research, National Zoological Park, Smithsonian Institution, Washington DC 20008, USA.

e-mail: rallsk@thegrid.net

1. Frankel, O. H. & Soule, M. E. *Conservation and Evolution* (Cambridge Univ. Press, 1981).
2. Saccheri, I. et al. *Nature* **392**, 491–494 (1998).
3. Ralls, K., Ballou, J. D. & Templeton, A. *Conserv. Biol.* **2**, 185–193 (1988).
4. Frankham, R. *Annu. Rev. Genet.* **29**, 305–327 (1995).
5. Frankham, R. *Conserv. Biol.* **9**, 792–799 (1995).
6. Lande, R. *Science* **241**, 1455–1460 (1988).
7. Caro, T. M. & Laurenson, M. K. *Science* **263**, 485–486 (1994).
8. Caughley, G. J. *Anim. Ecol.* **63**, 215–244 (1994).
9. Mills, L. S. & Smouse, P. E. *Am. Nat.* **144**, 412–431 (1994).
10. Frankham, R. *Heredity* **78**, 311–327 (1997).
11. Frankham, R. *Conserv. Biol.* (in the press).
12. Frankham, R. *Genet. Res.* **66**, 95–107 (1995).
13. Newman, D. & Pilon, D. *Evolution* **51**, 354–362 (1997).

Apoptosis

Phagocytic docking without shocking

John Savill

Apoptosis, a programmed form of cell death central to health and disease, is both topical and tidy. Doomed cells are swiftly identified and engulfed by phagocytes. But what are the mechanisms concerned? Answers on two fronts come in the papers on pages 501 and 505 of this issue by Wu and Horvitz¹ and Devitt et al.²

In normal tissues apoptosis was overlooked for many years because it is inconspicuous — intact dying cells are recognized, ingested and degraded beyond histological recognition by scavenger cells³. These can be neighbours acting as 'semi-professional' phagocytes or voracious experts of the macrophage line. Phagocyte recognition of 'apoptotic self' is also essential in protecting tissues from inflammatory injury due to leakage of noxious contents from dying cells⁴. During apoptosis these are safely packaged into membrane-bound bodies tagged with 'eat me' signals such as exposed phosphatidylserine⁵. However, there has been remarkably little study of how apoptotic cells reach their unmarked graves within phagocytes. Hence the value of

the new work, on the nematode worm, *Caenorhabditis elegans*¹, and human cells², respectively.

Caenorhabditis elegans must be a strong candidate for the 'most valued player' award in apoptosis research. Mutations affecting developmental cell deaths in this worm identify *ced* (cell death abnormal) genes which are homologous to, and sometimes interchangeable with, key elements in human apoptosis⁶. At least six genes regulate the semi-professional engulfment of cell corpses in *C. elegans*, but *ced-2*, *ced-5* and *ced-10* mutants also exhibit a fascinating and specific defect in migration of the distal tip cells of the gonad.

Wu and Horvitz¹ have now cloned and sequenced *ced-5*. They have discovered that this gene encodes a protein similar to DOCK180, a human cytoplasmic molecule which bears an SH3-domain 'passport' that allows it to interact with signalling pathways⁷. Interestingly, farnesylated DOCK180 can drive cell spreading, implying that it is involved in the regulation of cell movement by tyrosine kinases. Furthermore, the

authors¹ found that DOCK180 rescued distal-tip-cell migration in a *ced-5* mutant. It was perhaps too much to expect concomitant rescue of the engulfment defect; nevertheless, this was achieved by regulated expression of wild-type CED-5 protein, apparently through its effects on engulfing cells. Emboldened by the sequence similarities between CED-5, DOCK180 and the *Drosophila* protein Myoblast City, which is implicated in myoblast fusion⁸, Wu and Horvitz make the provocative suggestion that CED-5 is involved in the cytoskeletal reorganization required for an engulfing cell to extend its surface around a dying cell during phagocytosis.

Which phagocyte transmembrane molecules might trigger CED-5-mediated 'interment' of an apoptotic cell? The complete sequences of the five other *ced* genes involved in engulfment have yet to be published, but candidate molecules have been provided by a cell biological rather than a genetic approach. The first clues came from a simple *in vitro* model⁹, in which monosaccharides inhibited macrophage binding of apoptotic thymocytes, implicating phagocyte lectins such as the asialoglycoprotein receptor in recognition of sugar changes on apoptotic cells.

By backing hunches, other cell biologists used a similar inhibitor-based approach *in vitro* to identify further suspects — class A scavenger receptors¹⁰, a distinct scavenger receptor now identified as macroscialin¹¹, and ABC1, a mammalian ABC transporter¹², which may represent a homologue of the *C. elegans* engulfment gene *ced-7*. Two multi-functional adhesion molecules that bind bridging thrombospondin were also implicated in phagocyte recognition of apoptotic cells in culture. These were the $\alpha_v\beta_3$ vitronectin receptor integrin¹³ and the class B scavenger receptor CD36 (ref. 14), the index member of a family of phagocytic molecules that includes human CLA-1 and *Drosophila* Croquemort. Because $\alpha_v\beta_3$ and CD36 can signal via tyrosine kinases, these receptors could interact with CED-5-like molecules and the phagocyte cytoskeleton in promoting engulfment. Figure 1 shows a line-up of the molecules thought to be involved.

Could there be 'tethering' receptors which, by virtue of their high mobility within the phagocyte plasma membrane, capture apoptotic cells and shuttle them to the phagocytic machinery? One painstaking and systematic approach to this question involved screening a huge number of previously generated monoclonal antibodies in an assay of apoptotic lymphocyte tethering to macrophages. Most notably, the unclassified monoclonal antibody 61D3 blocked tethering and phagocytosis¹⁵.

In the second paper in this issue, Devitt et al.² now report expression cloning of the

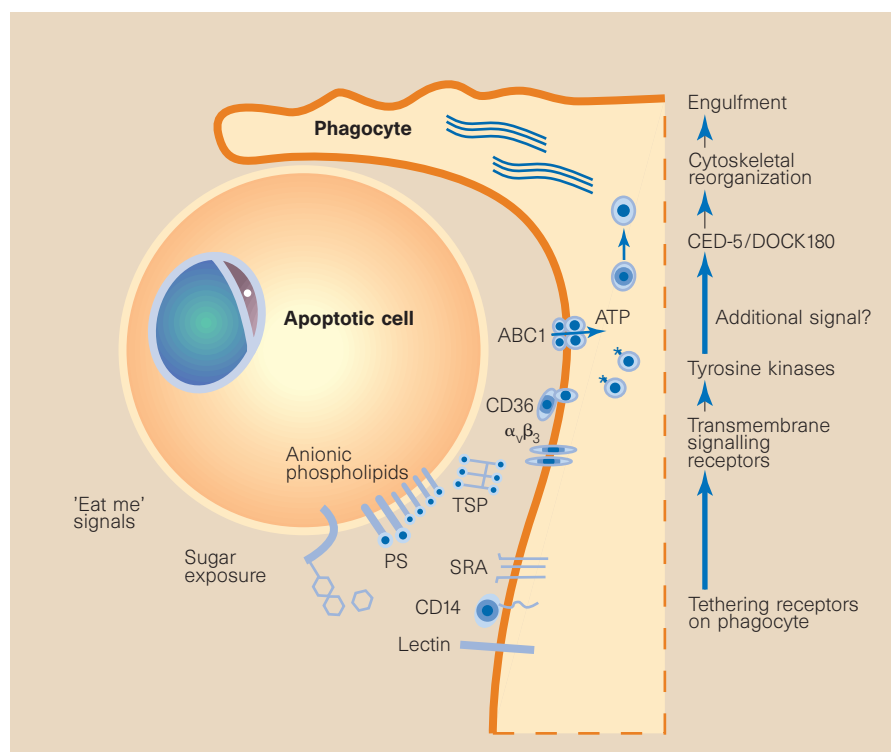


Figure 1 A model of phagocyte engulfment of an apoptotic cell, which follows phagocyte recognition that the cell is to be destroyed. The figure shows the molecules already implicated in these processes, and the newly proposed^{1,2} functions of CED-5 and CD14. TSP, bridging thrombospondin; SRA, class A scavenger receptor; PS, phosphatidylserine.

61D3 antigen. It turns out to be CD14, a glycosylphosphatidylinositol- (GPI-) anchored protein of relative molecular mass 55K, which binds the lipid A portion of bacterial lipopolysaccharide to trigger phagocyte activation *in vitro* and the septic-shock syndrome *in vivo*. This is surprising, as various reports demonstrate that uptake of large numbers of apoptotic cells does not stimulate¹⁶ and may even inhibit^{5,17} release of pro-inflammatory mediators from macrophages.

However, CD14 may bind certain lipids without triggering phagocyte activation¹⁸, including anionic phospholipids akin to those exposed by apoptotic cells¹⁹, and Devitt and colleagues' transfection of COS cells confirmed that CD14 is directly involved in tethering apoptotic cells. Finding out how this GPI-linked molecule engages the 'non-inflammatory' signalling pathways triggered by phagocytic transmembrane receptors will throw new light on the properties of CD14, extending the medical importance of this molecule beyond its known participation in septic shock. Indeed, further characterization of phagocyte recognition and processing of apoptotic cells²⁰ *in vivo* may reveal specific defects of clearance in persistent inflammatory and autoimmune diseases, such as some types of arthritis, and could lead to new treatments.

The new reports^{1,2} make an important contribution towards understanding a com-

plex process (Fig. 1) — one in which partially redundant phagocyte molecules may first tether apoptotic cells, then trigger transmembrane signalling which leads to activation of CED-5 homologues and closure of the phagocytic coffin around the moribund cell. □

John Savill is in the Division of Renal and Inflammatory Disease, University Hospital, Nottingham NG7 2UH, UK.
e-mail: john.savill@nottingham.ac.uk

1. Wu, Y.-C. & Horvitz, H. R. *Nature* **392**, 501–504 (1998).
2. Devitt, A. *et al. Nature* **392**, 505–508 (1998).
3. Wyllie, A. H. *Br. Med. Bull.* **53**, 451–465 (1997).
4. Savill, J. & Haslett, C. *Semin. Cell Biol.* **6**, 385–393 (1995).
5. Fadok, V. A. *et al. J. Clin. Invest.* **101**, 890–898 (1998).
6. Ellis, R. E., Yuan, J. & Horvitz, H. R. *Annu. Rev. Cell Biol.* **7**, 663–698 (1991).
7. Hasegawa, H. *et al. Mol. Cell. Biol.* **16**, 1770–1776 (1996).
8. Erickson, M. R. S., Galletta, B. J. & Abmayr, S. M. *J. Cell. Biol.* **138**, 589–603 (1997).
9. Duvall, E., Wyllie, A. H. & Morris, R. G. *Immunology* **56**, 351–358 (1985).
10. Platt, N. *et al. Proc. Natl Acad. Sci. USA* **93**, 12456–12460 (1996).
11. Ramprasada, M. *et al. Proc. Natl Acad. Sci. USA* **92**, 9580–9584 (1995).
12. Luciani, M. F. & Chimini, G. *EMBO J.* **15**, 226–235 (1996).
13. Savill, J. *et al. Nature* **343**, 170–173 (1990).
14. Ren, Y. *et al. J. Exp. Med.* **181**, 1857–1862 (1995).
15. Flora, P. K. & Gregory, C. D. *Eur. J. Immunol.* **24**, 2625–2632 (1994).
16. Meagher, L. C. *et al. J. Leuk. Biol.* **52**, 269–273 (1992).
17. Voll, R. E. *et al. Nature* **390**, 350–351 (1997).
18. Delude, R. L. *et al. Proc. Natl Acad. Sci. USA* **92**, 9288–9292 (1995).
19. Yu, B., Hailman, E. & Wright, S. D. *J. Clin. Invest.* **99**, 315–324 (1997).
20. Albert, M. L., Sauter, B. & Bhardwaj, N. *Nature* **392**, 86–89 (1998).

Daedalus

High radiation levels

Last week Daedalus was designing a free-flying helicopter rotor based on the Crookes radiometer principle. Each blade is black on one side to absorb sunlight, and bright on the other to reflect it. In sunlight, the black faces warm up; impinging air molecules rebound from them with added momentum; their reaction spins the rotor. This solar-powered 'altoradiometer' will fly very high in the atmosphere, where the pressure is low enough for this thermal transpiration process to exert a useful force. Sadly, it loses lift at nightfall, and may drift down too low to regain power and height the next day.

Daedalus now has an improvement. He plans to replace the black surfaces on the rotor blades with a radioactive coating. Nuclear decay will then keep the surfaces warm. The craft will be independent of the Sun, and will fly as long as its isotope layer lasts. Furthermore, the rotor will gain added thrust from the recoil of its energetic emissions. Such a radioactive coating will be highly dangerous, but Daedalus is undaunted. Once launched in the stratosphere, his craft will fly indefinitely at a safe height. Only when its radioactivity has safely decayed will it descend harmlessly to Earth.

At first Daedalus saw his craft as merely an unmanned high-altitude research vehicle. Its radioactivity would generate a local aurora; at night it would be visible as a little glow drifting across the sky, giving spectroscopic details of the air around it. But he now sees it as a splendid ecological development. Much of the world's radioactive waste could be lofted up into the stratosphere on altoradiometers, where it would wander around doing no harm. Even better, it could repair the ozone layer. Nuclear radiation can convert oxygen to ozone far more efficiently than solar ultraviolet light. A swarm of altoradiometers drifting about through the ozone layer, their huge rotor blades contacting vast swathes of air, could generate ozone faster than it could be destroyed by halocarbon radicals wafting up from below. If their radioactive coatings were blackened to absorb sunlight, the resulting added oblique thrust during the hours of daylight would even tend to push them towards the poles, where the ozone is most depleted. Dedicated ecologists will be outraged at the thought of one wicked industrial menace being safely disposed of by countering another; but the rest of us will be well content.

David Jones

Obituary

Haroun Tazieff (1914–98)

Volcanologist, and authority on natural hazards

At a time when science is becoming increasingly arcane and remote from human experience, we feel a growing need for those few people like Haroun Tazieff who can still convey to the general public an appreciation of the marvels of the natural world. Tazieff died at home in Paris on 2 February, aged 83.

Born in Warsaw of a Russian father who perished in the early days of the First World War, Tazieff and his mother moved first to Saint Petersburg and then, with the outbreak of the revolution, to Belgium, where he grew up in very modest circumstances. After serving in the Belgian army and resistance during the Second World War, he completed his studies in geology and agronomy and went to the Belgian Congo to work as a geologist and mining engineer. While there, he wrote several scientific papers on the alkaline igneous rocks of the African volcanoes, but after experiencing the 1948 eruption of Kituro he became completely devoted to studies of active volcanism. His first book, *Cratères en feu*, published in 1951 (and in English the following year as *Craters of Fire*), was an immediate popular success thanks to the sense of adventure and grandeur of natural forces his writing conveyed. In the following years he produced a total of 23 books and six films of extraordinary artistry and popular appeal.

After returning to Belgium, Tazieff taught at the University of Brussels and later at the Universities of Paris and Orsay. In 1957 he set up the Centre de Volcanologie in Belgium and in 1961 organized the International Institute of Volcanology at Catania, Sicily. He was director of the volcanological laboratory at the Institut de Physique du Globe in Paris and later established a laboratory under the French national research agency CNRS at Gif-sur-Yvette. His ability to sample erupting lavas close to their source resulted in some of the best gas analyses and temperature measurements ever obtained. Some of the instruments he and his team designed for this work are now widely used in science and industry. But although his interests mainly centred on understanding eruptive mechanisms, he always maintained a primary concern for the human aspects of natural hazards.

Tazieff served as an expert on volcanic hazards around the world, and was an adviser on natural hazards for several



regions in France. In 1976 he was called to the island of Guadeloupe to assess the potential hazards of an eruption from the Soufrière volcano. When the international team recommended evacuating all inhabitants from the slopes of the volcano, Tazieff said that this was a costly over-reaction and that the volcano posed no immediate danger. In the end, 73,600 people were evacuated at great expense, and the eruption ended quite peacefully. In a brief article (*Nature* 269, 96–97; 1977), Tazieff summarized his views and proposed a code of professional conduct for volcanologists charged with assessing the risks posed by eruptions in populated regions.

At that time, I was editor of the *Journal of Volcanology and Geothermal Research*. Convinced that the professional questions Tazieff raised deserved the serious attention of volcanologists, I wrote a short note summarizing his main points (*J. Volcanol. Geotherm. Res.* 4, 1; 1978). So as not to portray Tazieff's views as those of the journal, I used the nom de plume 'Derek Bostok'. The note brought an immediate response, mostly negative. The main argument was that the evacuation was justified, because volcanologists could not foresee all possible courses that the eruption could take. The consequences of a major eruption would have been too great to justify taking any risk, no matter how small it might seem. Despite the heated nature of the arguments, the 'Bostok Affair', as it came to be known, provided a healthy appraisal of the role of geologists in dealing with natural hazards.

In the ensuing years, Tazieff continued to contribute provocative ideas which can be summarized in terms of three principles.

First, the responsibility for interpreting hazards must be left to specialists. The complexities of volcanism and the difficulty of anticipating the behaviour of individual volcanoes are not generally

appreciated by the geologist who has not had the opportunity to deal with a wide range of volcanic phenomena.

Second, the primary task is not to predict volcanic eruptions but rather to foresee catastrophic events. Few eruptions pose serious threats, but one must always consider the possibility of unexpected events. Even the most experienced observer cannot anticipate every possible development in a large mature volcano, especially if it has no historical record of activity.

Third, and possibly most important, the role of the professional volcanologist is that of adviser to public officials who must decide on an appropriate response to his advice. Self-appointed experts who voice poorly informed opinions through the press have caused needless panic and costly over-reaction. Even the best experts may have conflicting opinions — if these are voiced publicly, responsible officials lose confidence and resort to the most conservative measures, even when these entail great expense and hardship.

Tazieff increasingly concerned himself with all types of natural hazards and environmental problems. Between 1984 and 1986 he held the post of Secretary of State for Prevention of Natural and Technological Disasters under Laurent Fabius, then Prime Minister of France. He continued to serve as adviser at various governmental levels until only a few years before his death.

It never seemed to bother Tazieff that he was never accepted as a member of the 'club'; he clearly preferred the role of iconoclast and with time became increasingly outspoken. His last published work, *J'accuse une dernière fois* (*Préventique* No. 37, 4–15; 1998), was a broadside against 'opportunists' of all kinds, especially politicians. He also expressed contempt for scientists who solicit funds for research with the ostensible aim of alleviating natural hazards but who in fact use them to pursue esoteric research.

It is not surprising that combative remarks such as these raised strong reactions on the part of 'serious' scientists. In the end, however, Tazieff achieved the two primary goals of his long, eventful career — he made scientists more conscious of their responsibilities to society, and he inspired a generation of ordinary citizens with an appreciation of the magnificent manifestations of nature.

Alexander McBirney

Alexander McBirney is in the Department of Geological Sciences, University of Oregon, Eugene, Oregon 97403, USA.
e-mail: mcbirney@darkwing.uoregon.edu

Saenredam's shapes

Painting 'portraits' of churches might seem a limiting pursuit. But the seventeenth-century Dutch artist Pieter Saenredam turned it into a paean of praise for the geometrical way in which we perceive space.

Martin Kemp

One of the peculiarities of linear perspective is that, the more rigorously we follow the principles of mathematical projection to create convincing illusions of forms in space, the more we become aware of their strange shapes on the flat plane of the picture surface. This is particularly true for geometrical forms that lie near the margins of wide or tall pictorial fields.

Those perspectivists who have most assiduously followed the geometrical rules have also exhibited the greatest awareness of the implications of the surface configuration of projected forms. Piero della Francesca in the Renaissance and Pieter Saenredam in seventeenth-century Holland stand supreme in this respect.

Saenredam specialized in paintings of churches. He transformed the apparently limited subject of sacred interiors into a genre that 'revealed' the geometrical art of seeing space. We know from his drawings and related notes that he made on-the-spot sketches and obtained accurate measurements from building surveys. From these he developed elaborate construction drawings, projecting measurements from a scale along the base of the drawing on to the orthogonals (the lines that converge on the vanishing point).

He was at pains to emphasize that his was an art of the geometrical observer. He consistently marked what he called the "eye point" on his first sketches, and typically inscribed his presence as a witness somewhere on the fabric of the building in the painted versions.

On the leftmost pier in the large 'portrait' of St Bavo's we read, "This is the cathedral Great Church of Haarlem in Holland. Pieter Saenredam finished painting this on February 27, 1648".

However, a close comparison of the final painting with a precise projection from his viewpoint shows that he has judiciously adjusted some of the shapes, most notably by stretching the two nearest arches upwards, to make his composition function to best effect on the surface of his wooden panel. He balances the requirements of precise illusion with the need to orchestrate the fragile network of lines that endow his surface composition with such tensile beauty.

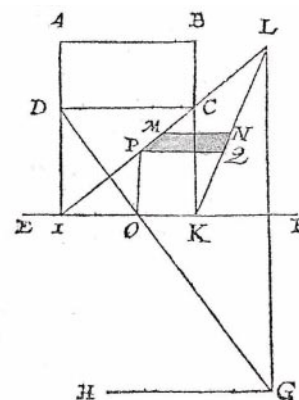
There is little doubt that Saenredam was fully alert to the way that projective geometry in his era was exploring the transformation and homologies of such figures as squares and triangles in perspective. He is



Saenredam's *Interior of St Bavo's Church, Haarlem*, 1648, National Gallery of Scotland, Edinburgh. *Perspective Projection of a Rectangle (right) from Stevin's book on perspective, 1605.*

known to have owned volumes by Simon Stevin, the mathematician, physicist and engineer. Stevin's treatise on perspective follows the uncompromisingly geometrical approach of Guidobaldo del Monte, whose perspective book of 1600 set the standard. A series of abstract exercises shows how to project geometrical shapes positioned at any angle behind a plane from a viewpoint at any given height and distance.

For Guidobaldo and Stevin, and for the great French master of projective geometry, Girard Desargues, the abstract mathematics of perspective (for all its practical application in art) held a cerebral fascination sufficient unto itself. For Saenredam, it was the sensate geometry of light that provided the



beginning and end of his quest to reveal the visual wonders of divine architecture. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

First protozoa-trapping plant found

Some 450 species of insectivorous plants have been discovered since Darwin published his work on the subject in 1875¹. We report here the first known case of a plant that traps and digests protozoa.

Insectivorous plants colonize nutrient-poor places, and their major source of phosphorus, nitrogen and other elements is trapped and digested insects^{2,3}. Extravagant trap-like structures are seen in the genus *Genlisea* (a member of the Lentibulariaceae, with the carnivorous genera *Pinguicula* and *Utricularia*) which occurs mainly in nutrient-poor white sands and moist rock outcrops in South America and tropical Africa. However, proof of the plant's carnivorous nature was previously lacking.

Genlisea species are rare in the wild and difficult to cultivate. Most form a small rosette about three centimetres in diameter, close to the ground, with linear or spatulate leaves. The yellow or violet flowers, similar to those of the related *Antirrhinum*, are borne on an inflorescence some 20 cm high.

When the rosette of *Genlisea* is dug up, pale bundles of root-like organs up to 15 cm long are revealed (Fig. 1a). However, the plant is rootless and these organs are subterranean leaves, highly modified and lacking chlorophyll⁴⁻⁶. The long, solid, basal part of the leaf attached to the rhizome opens into a hollow utricle. This contracts into a narrow tubular channel (the 'neck piece') which divides into two arms at an angle of 90–130° so that the apical part of the leaf is shaped like an inverted letter 'Y'. The helically contorted arms are hollow, with an average inside diameter of 200 micrometres. The arms bear slit-like openings (width 400 µm, height 180 µm) lined with rows of hairs pointing towards the utricle, and numerous glands.

Since Darwin's time, it has been postulated that these specialized leaves are traps for catching prey, but there has been no proof of carnivory. Arthropod remnants have only exceptionally been found in the traps. This, and the dimensions of the traps, suggested to us that the subterranean leaves might function as highly

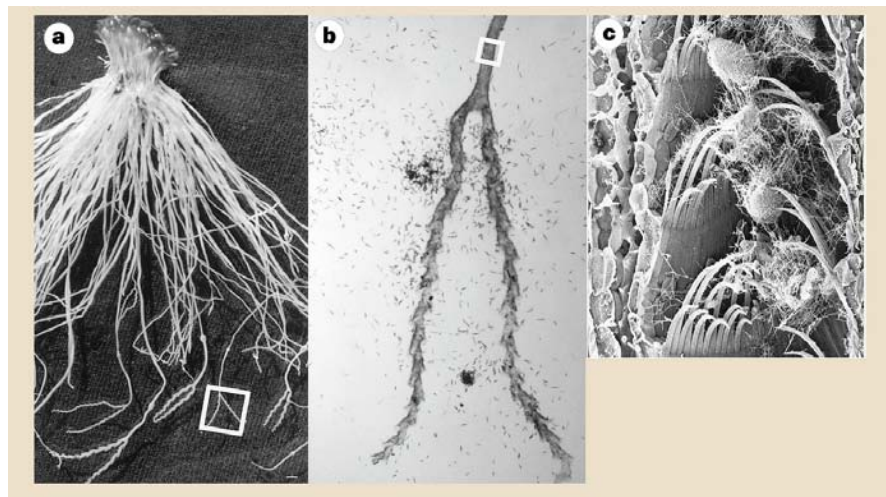


Figure 1 *Genlisea aurea* protozoan traps. **a**, Green rosette of assimilating leaves (on top) and a large bundle of root-like subterranean leaves acting as traps. **b**, A single *Genlisea* trap with protozoa attracted and concentrated around the utricle and the tube. **c**, Scanning electron micrograph of the tube after a 'feeding' experiment with *Paramecium caudatum*, fixed with osmium tetroxide. Captured protozoa are visible. The rows of long hairs prevent escape.

specialized traps for catching protozoa.

To test this idea, we cultivated species of *Genlisea* in a greenhouse. Laboratory experiments were conducted with *G. aurea*, *G. violacea* and *G. margaretae*. Intact plants were placed in a Petri dish and ciliates such as *Blepharisma americana* were added.

The root-like subterranean organs of *Genlisea* proved attractive to protozoa (Fig. 1b). A few minutes after starting the experiment, numerous protozoa had already entered the traps (Fig. 1c). In contrast, living roots from plants occurring in the same habitats — such as *Eriocaulon plumale* — failed to attract ciliates.

The presence of acid phosphatases and esterases in *G. africana*⁷ suggests that the ciliates are digested. Additional experiments proved the existence of an attractant and indicated that *Genlisea* attracts protozoa chemotactically, trapping them in its subterranean leaves.

Genlisea can therefore be regarded as a highly specialized protozoan trap. Experiments using ciliates marked with the isotope sulphur-35 demonstrated its uptake by

the trap leaves of *Genlisea*. After two days, ³⁵S was traceable in the rosette leaves.

Field investigations at a natural site in the northern part of the Ivory Coast in west Africa indicate that, at this site at least, nine different ciliate species, as well as other protozoa, are trapped in large amounts by *G. stapfii*. Thus, 125 years after Charles Darwin's initial postulations, the puzzle of *Genlisea*'s feeding habits seems to have been solved.

Wilhelm Barthlott, Stefan Porembski, Eberhard Fischer, Björn Gemmel

Botanisches Institut der Universität Bonn,
Meckenheimer Allee 170, D-53115 Bonn, Germany
e-mail: barthlott@uni-bonn.de

1. Darwin, C. *Insectivorous Plants* (Murray, London, 1875).
2. Juniper, B. E., Robins, R. J. & Joel, D. M. *The Carnivorous Plants* (Academic, London, 1989).
3. Lloyd, F. E. *The Carnivorous Plants* (Chronica Botanica, Waltham, 1942).
4. Goebel, K. *Organographie der Pflanzen* part 1, 3rd edn (Fischer, Jena, 1928).
5. Warming, E. *Medd. Kjob.* 1, 33–70 (1874).
6. Reut, M. S. *Bot. Helv.* 103, 101–111 (1993).
7. Heslop-Harrison, Y. in *Lysosomes in Biology and Pathology 4* (eds Dingle, J. T. & Dean, R. T.) 525–578 (North Holland, Amsterdam, 1975).

Domains of rasGAP and rhoGAP are related

Guanine-nucleotide-binding (G) proteins are 'switched off' by the hydrolysis of their bound GTP. The GTP bound to the G protein Ras is hydrolysed intrinsically at a very slow rate¹, so, *in vivo*, Ras is turned off by GTPase-activating proteins (GAPs). The

structures of GTPase-activating domains (GAP domains), which usually occur as parts of larger proteins, indicate that the Ras and Rho families of small G proteins and their GAPs evolved in parallel.

Determination of the structure of the GAP domain from p120GAP, first in isolation² and then in a Ras-rasGAP complex³, showed that an arginine residue from the GAP domain plays a crucial role in catalysis^{1,3}. GAPs for the Rho family have also

been characterized and the structure of a GAP domain from p50rhoGAP has been determined in isolation⁴ and in a transition-state complex with RhoA⁵. The rhoGAP domain provides a crucial arginine residue to the active site of RhoA.

A domain homologous to rhoGAPs but lacking GAP activity is found in the regulatory p85 subunit of phosphatidylinositol-3-OH kinase. The structure of this rhoGAP-like domain from p85α, called

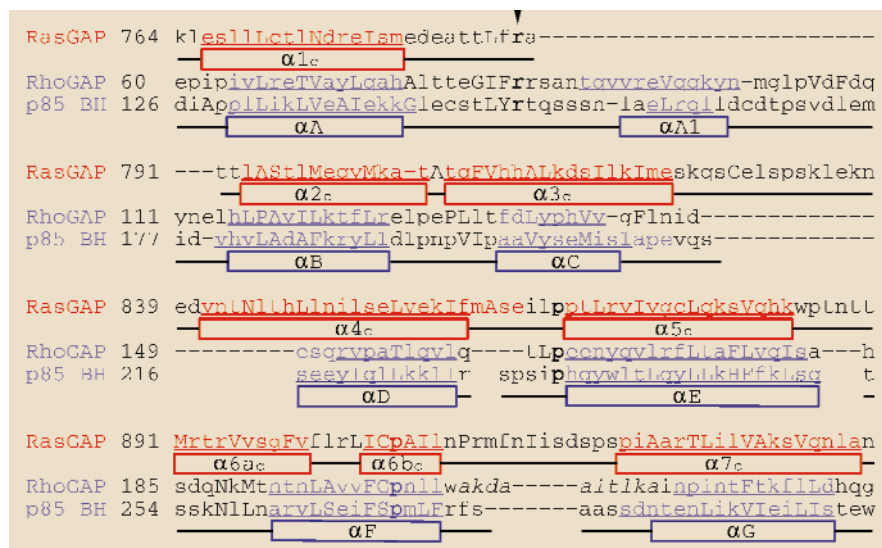


Figure 1 A structurally based sequence alignment of comparable regions of the rasGAP domain from p120-GAP (PDB code: 1wer), the rhoGAP domain of p50rhoGAP (PDB code: 1rgp) and the BH rhoGAP-like domain of the p85 subunit of phosphatidylinositol-3-OH kinase (PDB code: 1pbw). The computer program COMPARE²⁸ provided the alignment. α -Helices are underlined and shown in red (rasGAP) or blue (rhoGAP, p85 BH); 3_{10} helices are not underlined. The nomenclature for helices is as originally defined^{26,4}. Residues labelled in upper-case lettering are inaccessible to solvent. The position of Arg 789 (rasGAP)/Arg 85 (rhoGAP)/Arg 151 (p85 BH) is indicated by the arrow. Residues between helix α F and α G are disordered in the rhoGAP structure (in italics).

p85 BH (where BH indicates a BCR-homology domain), has also been determined⁶. The release of the coordinates for the rhoGAP and rasGAP domains allowed us to compare their structures. We found that the rhoGAP and rasGAP domains are clearly related (my Fig. 1 and Fig. 1 in the letter below by Rittinger *et al.*) and must have derived from the same ancestral protein.

To see the equivalence of the structures more clearly (in Fig. 1 of Rittinger *et al.*), place your finger on the top of helix F (of rhoGAP) and imagine pushing it away from you and down so that it lies more nearly parallel to helix G and between helices G and E (helices E and A then move to the left). The rasGAP and rhoGAP sequences (in my Fig. 1) are only 6% identical, whereas rasGAP and p85 BH are 13% identical (the two rhoGAP-like domains, whose similarity was identified at the sequence level, have a 17% sequence identity).

In the transition-state-analogue complexes of both Ras-rasGAP³ and Rho-rhoGAP⁵, helix α 6/ α F interacts with the switch II loop on Ras/Rho. This interaction helps to stabilize the conformation of the switch II loop which provides residue Gln 61 (Gln 63 in Rho) for the active site. The large structural shift in the relative orientation of helix α 6/ α F of rasGAP/rhoGAP may have occurred to compensate for the different structures of the switch II loop in Ras and Rho.

This equivalence in structure points to a parallel evolution by Ras and Rho G proteins and their GAPs.

Benjamin Bax

Department of Crystallography, Birkbeck College,
Malet Street, London WC1E 7HX, UK
e-mail: b.bax@mail.cryst.bbk.ac.uk

1. Ahmadian, M. R., Stege, P., Scheffzek, K. & Wittinghofer, A. *Nature Struct. Biol.* **4**, 686–689 (1997).
2. Scheffzek, K., Lautwein, A., Kabsch, W., Ahmadian, M. R. & Wittinghofer, A. *Nature* **384**, 591–596 (1996).
3. Scheffzek, K. *et al. Science* **277**, 333–338. (1997).
4. Barrett, T. *et al. Nature* **385**, 458–461 (1997).
5. Rittinger, K., Walker, P. A., Eccleston, J. F., Smerdon, S. J. & Gamblin, S. J. *Nature* **389**, 758–762 (1997).
6. Mussachio, A., Cantley, L. C. & Harrison, S. C. *Proc. Natl Acad. Sci. USA* **93**, 14373–14378 (1996).
7. Sali, A., & Blundell, T. L. *J. Mol. Biol.* **212**, 403–428 (1990).
8. Zhu, Z.-Y., Sali, A. & Blundell, T. L. *Protein Engng* **5**, 43–51 (1992).

Support for shared ancestry of GAPs

The Ras superfamily of monomeric GTPase proteins comprises six subfamilies, each with specific cellular functions, but all these subfamilies share a common fold and a common mechanism for hydrolysing GTP. Unlike their cognate guanine-nucleotide-binding (G) proteins, GTPase-activating proteins (GAPs) specific for the Rho or Ras subfamilies show no significant homology at the level of protein sequence. However, we have now noticed a structural similarity which suggests that, like their G proteins, these GAPs have evolved from a common ancestor.

Small G proteins belonging to the Ras superfamily act as signalling molecules con-

trolling a wide range of biological processes. Crucial to their signalling function is the conformational switching that takes place between their active, GTP-bound and inactive, GDP-bound forms¹. They have a slow, intrinsic rate of GTP hydrolysis which is substantially accelerated by GAPs specific to each subfamily. Crystallographic coordinates for rhoGAP and rasGAP are now available^{2,3} and we have compared them using the program LSQMAN (ref. 4).

Superposition of the two GAPs shows that they share a core structure made up of seven α -helices (Fig. 1). The α -helices making up this core pack against each other in a related, but not identical, fashion. However, the automated alignment puts the catalytic arginine residues (R85 in rhoGAP and R789 in rasGAP) into approximately the same position and, by implication, locates the active sites of the bound G proteins.

It is particularly interesting to note that in both structures the catalytic arginine residue is located on a surface loop and is preceded by two bulky, hydrophobic residues (isoleucine and phenylalanine at residues 83 and 84, respectively, in rhoGAP, and leucine and phenylalanine at positions 787 and 788, respectively, in rasGAP). These residues are held in a hydrophobic clamp made up of residues from the B(2c), E(5c) and F(6c) helices that act to anchor the catalytic arginine loop to the core of the domain.

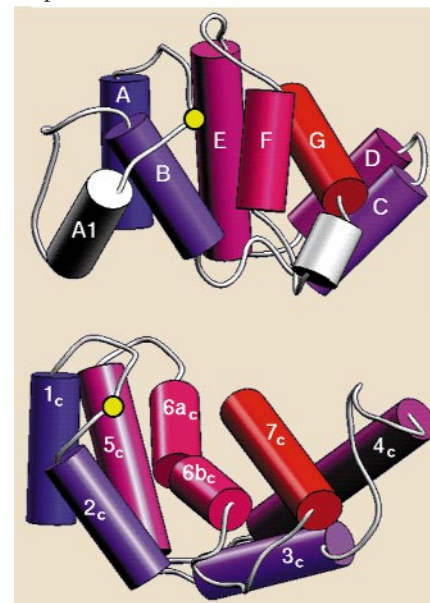


Figure 1 The common core structures of the catalytic domains of p50rhoGAP (top) and p120rasGAP (bottom), as aligned by the computer program LSQMAN, in helical-tube representation. Helical segments forming the core structure (helices A-G and 1c to 7c in rhoGAP and rasGAP, respectively) are coloured from blue at their amino termini to red at their carboxy termini. Helical insertions that do not form part of the core motif are grey. The positions of the catalytic arginines R85(rhoGAP) and R789(rasGAP) are indicated by yellow circles.

The published structures of the transition-state-analogue complexes of RhoGDP·AlF₄/p50rhoGAP (ref. 5) and RasGDP·AlF₃/p120rasGAP (ref. 6) show that the two GAP domains bind their respective G proteins in related but distinct ways. rhoGAP binds to Rho primarily through a surface made up by the helices B and F, whereas rasGAP binds to Ras through the antiparallel helices 6c and 7c.

These differences in G-protein-binding surfaces arise because of the different spatial relationship between helices A–B and E–F of rhoGAP and 1c–2c and 5c–6c of rasGAP. Relative to the rhoGAP helices E and F, the carboxy terminal (or top half) of 5c bends back into the plane of the page while the whole of 6c rotates further in a similar direction.

Nonetheless, both GAP domains present their catalytic arginine residues to the active sites of the G protein in very similar ways, which is a consequence of the equivalence in function of the arginine loop in these two systems. It will be intriguing to see whether other GAPs associated with the Ras superfamily contain the same core structure, obey the same topology, and make similar use of hydrophobic clamping to orient the catalytic arginine loop.

Katrin Rittinger, William R. Taylor, Stephen J. Smerdon, Steven J. Gamblin
Structural Biology Group,
National Institute for Medical Research,
The Ridgeway, London NW7 1AA, UK
e-mail: s-smerdo@anika.nimr.mrc.ac.uk

1. Boguski, M. S. & McCormick, F. *Nature* **366**, 643–654 (1993).
2. Barrett, T. *et al.* *Nature* **385**, 458–461 (1997).
3. Scheffzek, K., Lautwein, A., Kabsch, W., Ahmadian, M. R. & Wittlinghofer, A. *Nature* **384**, 591–596 (1996).
4. Kleywegt, G. J. & Jones, T. A. *Meth. Enzymol.* **277**, 525–546 (1997).
5. Rittinger, K., Walker, P. A., Eccleston, J. F., Smerdon, S. J. & Gamblin, S. J. *Nature* **389**, 758–762 (1997).
6. Scheffzek, K. *et al.* *Science* **277**, 333–338 (1997).

Meningitis bacterium is viable without endotoxin

The outer membrane of Gram-negative bacteria contains lipopolysaccharide (LPS) as its outer monolayer. This is anchored to the membrane by lipid A, which is responsible for LPS's activity as an endotoxin¹. In *Escherichia coli*, conditionally lethal mutants have been reported for the genes involved in the early steps of lipid A biosynthesis², suggesting that this part of the LPS molecule is essential for bacterial growth. However, we have isolated a mutant of *Neisseria meningitidis* which is viable in spite of an early block in lipid A biosynthesis that causes a loss of endotoxin activity.

The protein LpxA is responsible for adding the O-linked 3-OH fatty acid to

UDP-N-acetylglucosamine, which is the first committed step in the lipid A biosynthesis pathway^{3,4}. The *E. coli* and *N. meningitidis* LpxA proteins differ in their specificity, preferring 3-OH C₁₄ and 3-OH C₁₂ fatty acyl chains, respectively^{5,6}.

We constructed a hybrid *lpxA* gene in which the meningococcal N-terminal part was replaced by the corresponding part of *E. coli lpxA*. Plasmid pHBK30, carrying this hybrid gene and an upstream kanamycin-resistance cassette, was used for allelic replacement of the wild-type *lpxA* gene on the chromosome of meningococcal strain H44/76. Lipopolysaccharide of the H44/76[pHBK30] mutant and the wild-type strain was compared by Tricine-SDS-PAGE followed by silver staining for carbohydrates (Fig. 1a). No LPS could be detected in the hybrid derivative by this method, even when higher amounts of cell lysates were loaded on the gel.

A panel of LPS and monoclonal antibodies specific for outer-membrane proteins (OMPs) was tested in a whole-cell enzyme-linked immunosorbent assay (ELISA)^{7,8}. The mutant strain did not bind any of the LPS-specific antibodies (either immunotype-specific or broadly crossreactive), whereas the OMP-specific antibodies showed similar binding patterns for mutant and wild type. This apparent OMP similarity was confirmed when outer-membrane complexes (OMCs) of H44/76[pHBK30] and H44/76 were isolated by sarkosyl extraction and analysed by SDS-PAGE (Fig. 1b). Both strains show equal amounts of the class 1, 3 and 4 principal OMPs.

As LPS of H44/76[pHBK30] could not be detected, it was not clear whether it was present at all. Therefore, the mutant and wild-type strain were tested in a chromogenic *Limulus* (LAL) assay⁹ for endotoxin, using the QCL-1000 kit from BioWhittaker. The results of the LAL assay on cell suspensions showed no significant endotoxin activity for H44/76[pHBK30] over meningococcal medium (0.3 and 1.7 endotoxin units per millilitre, respectively), in contrast to 21.7 × 10⁴ endotoxin units per millilitre for the wild type.

Although H44/76[pHBK30] is viable, it had a reduced growth rate. When grown overnight on GC agar plates, the mutant strain produced much smaller colonies. In liquid medium, the doubling time during exponential growth was about 50% longer than in the wild-type strain H44/76. The morphology of H44/76[pHBK30] and its parent strain was examined by electron microscopy of ultrathin sections (Fig. 1c). The ultrastructure of the outer membrane could be clearly discerned in the LPS-deficient mutant and was not visibly altered.

Strain H44/76[pHBK30] is viable and possesses an outer membrane, despite lacking detectable LPS. It seems likely that the

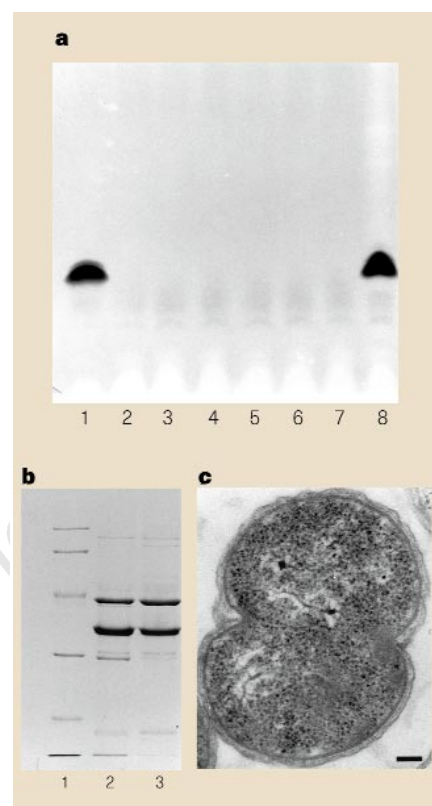


Figure 1 Analysis of *Neisseria meningitidis* outer membrane. **a**, Silver-stained Tricine-SDS-PAGE LPS gel of proteinase K-treated whole-cell lysates of H44/76 wild type (lanes 1 and 8) and six independent kanamycin-resistant transformants with pHBK30 (lanes 2–7). **b**, SDS-PAGE of outer membrane proteins from H44/76[pHBK30] (lane 2) and H44/76 wild type (lane 3); lane 1 contains relative molecular mass markers of 94K, 67K, 43K, 30K, 20.1K and 14.4K. **c**, Electron micrograph of a thin section of H44/76[pHBK30], showing normal appearance of the cell envelope with outer membrane, peptidoglycan layer within the periplasmic space, and inner membrane. Scale bar, 100 nm.

hybrid *lpxA* gene is inactive, either because of disrupted transcription/translation in our construct, or otherwise because the hybrid protein produced lacks enzymatic activity. We constructed a *lpxA*-knockout mutant by inserting a kanamycin-resistance cassette into the *Bst*EII site located at position 293 in the *lpxA* gene of plasmid pLA21 (a pUC18 derivative with a 2.1 kilobase *lpxD-fabZ-lpxA* insert). The resulting plasmid pLAK33 was linearized and transformed to strain H44/76, with selection for kanamycin resistance. The resulting colonies showed the same growth properties as the H44/76[pHBK30] mutant. In whole-cell ELISA, the *lpxA*-knockout mutant did not bind any of the LPS-specific monoclonal antibodies.

The absence of LPS was further confirmed by gas chromatography/mass spectrometry analysis of fatty acids present in OMC preparations, which showed that the lipid A-specific 3-OH C₁₂, whose addition

to UDP-*N*-acetylglucosamine is catalysed by LpxA, was present only in trace amounts in the mutant. Also, no other *O*-linked 3-OH fatty acids in the range C₁₀ to C₂₀ could be detected. LPS biosynthesis could be restored in this knockout mutant by retransformation with wild-type *lpxA*, showing that the observed LPS deficiency must result from mutation in this gene only.

The availability of LPS-deficient mutants will allow new approaches to vaccine development against *N. meningitidis*¹⁰ and the closely related pathogen *N. gonorrhoeae*, or any other bacteria for which such mutants can be isolated. Using an *lpxA* mutant, it will be much easier to purify OMPs or other cell-surface components without contamination by endotoxin. Also, the role of LPS in outer-membrane-vesicle or whole-cell vaccines (for example as adjuvant¹¹) can be investigated: it may be possible to substitute a less toxic compound. And at a more basic level, such mutants can be used for studying how LPS contributes to the structure and biogenesis of the outer membrane.

**Liana Steeghs, Ronald den Hartog,
Arie den Boer, Bert Zomer, Paul Roholl,
Peter van der Ley**

National Institute of Public Health
and the Environment, RIVM, PO Box 1,
3720 BA Bilthoven, The Netherlands
e-mail: peter.van.der.leij@rivm.nl

1. Zähringer, U., Lindner, B. & Rietschel, E. Th. in *Advances in Carbohydrate Chemistry and Biochemistry*, vol. 50 (ed. Horton, D.), 211–276 (Academic, San Diego, 1994).
2. Raetz, C. R. H. *Annu. Rev. Biochem.* 59, 129–170 (1990).
3. Anderson, M. S. & Raetz, C. R. H. *J. Biol. Chem.* 262, 5159–5169 (1987).
4. Coleman, J. & Raetz, C. R. H. *J. Bacteriol.* 170, 1268–1274 (1988).
5. Steeghs, L., Jennings, M. P., Poolman, J. T. & van der Ley, P. *Gene* 190, 263–270 (1997).
6. Odegaard, T. J. et al. *J. Biol. Chem.* 272, 19688–19696 (1997).
7. van der Ley, P., van der Biezen, J., Hohenstein, P., Peeters, C. & Poolman, J. T. *Infect. Immun.* 61, 4217–4224 (1993).
8. van der Ley, P. et al. *Mol. Microbiol.* 19, 1117–1125 (1996).
9. Friberger, P., Knos, M. & Mellstam, L. in *Endotoxins and their Detection with the Limulus Amebocyte Lysate Test* (eds Watson, S. W., Levin, J. & Novitsky, T. J.) 195–206 (Liss, New York, 1982).
10. Poolman, J. T. *Infect. Agents Dis.* 4, 13–28 (1995).
11. Nakano, M. & Matsuura, M. in *The Theory and Practical Application of Adjuvants* (ed. Stewart-Tull, D. E. S.) 315–335 (Wiley, Chichester, 1995).

Speed perception fogs up as visibility drops

Many horrendous vehicle accidents occur in foggy weather. Drivers know they should slow down because fog reduces visibility, but many still drive too quickly¹. The ‘blame’ for many such accidents may be due to a perceptual quirk: it appears that drivers think they are driving far more slowly than they actually are in foggy conditions, and therefore increase their speed.

We used a virtual-environment driving

simulator to show that, as fog increases and therefore reduces the contrast of the driver’s image, the apparent speed of the vehicle slows. Participants asked to drive at a certain speed drove faster as the scene became foggy.

An accurate representation of the current speed is normally provided to drivers by the speedometer. However, reading this instrument requires drivers to divert their gaze and attention from the road to the appropriate dial. In conditions of reduced visibility produced by fog, drivers are reluctant to divert their gaze from the road to the speedometer for fear of missing an object emerging from the fog². Hence it is exactly in conditions of reduced visibility caused by fog that drivers rely on their own perceptual judgement of speed.

Thompson³ has reported that the perceived speed of a moving-grating pattern depends on its level of contrast. As contrast decreases, the grating appears to drift more slowly. This effect has been replicated several times (and extended to other stimulus patterns⁴), but others have failed to replicate it or have shown that perceived speed is unaffected by random variations in contrast⁵. So it remains an open question as to whether this illusion of reduced speed with decreasing image-contrast will occur in conditions akin to those encountered when driving in fog.

We tested for the perceived slowing of a visual scene by conducting two experiments in a virtual environment that simulated the view from a vehicle moving along a road. Such a set-up allowed us to manipulate the visual stimulus in a highly specific manner.

Our first experiment involved showing the observer two scenes. In one scene, the weather was ‘clear’; in the other, it was either ‘clear’, ‘misty’ or ‘foggy’. The physical speed at which the two scenes appeared to move at the same speed was calculated (Fig. 1a) using standard psychophysical techniques. Subjects perceived the foggy scenes to be moving more slowly⁶.

Does this perceptual change affect driving speed in a more realistic task? We trained subjects (in clear conditions) to ‘drive’ a simulated vehicle at set speeds along a winding road, using a brake, accelerator and steering column.

In the experimental phase, the subject was given a target speed and told to adjust the car speed (while steering along the road) to this target, using the accelerator and brake. Subjects accelerated and decelerated as they thought appropriate for as long as they pleased, and then signalled that they believed that they were travelling at the target speed. Trials in which the weather was ‘clear’, ‘misty’ and ‘foggy’ were randomly interleaved. As the

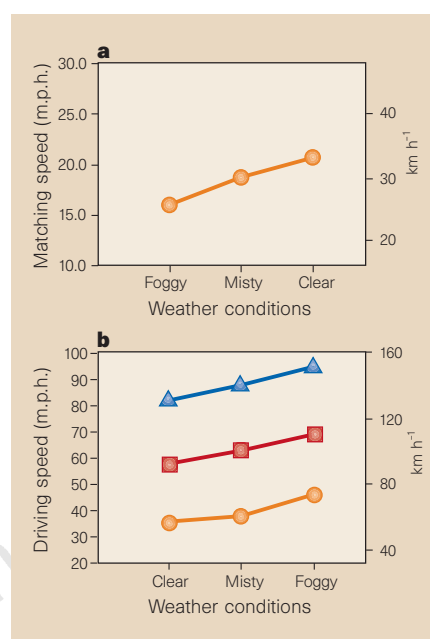


Figure 1 Sense of speed decreases in fog. **a**, Perceived matching speed averaged over all subjects ($n = 5$) in three weather conditions under passive viewing. Foggy scenes gave the impression of slower movement (one-way analysis of variance; $P < 0.001$). Simulation was performed using a Silicon Graphics Crimson Reality Engine that updated the viewing screen (60 Hz VDU monitor) at a rate of 15 Hz. ‘Fog’ was simulated by blending a partially transparent polygon over each pixel. For each pixel, the red, green and blue guns levels were recalculated (R') from the original pixel values (R) drawn from memory according to the formula, $R' = 255F + (1 - F)R$. ‘Clear’ conditions: $F = 0$, contrast (root mean square) = 0.61; ‘misty’: $F = 0.5$, contrast = 0.20; ‘foggy’: $F = 0.8$, contrast = 0.07. **b**, The actual speed driven under three weather conditions ($n = 9$). Target speeds: circles, 30 m.p.h. (48 km h⁻¹); squares, 50 m.p.h. (80 km h⁻¹); triangles, 70 m.p.h. (112 km h⁻¹). Analysis of variance showed significant effects of target speed ($P < 0.0001$) and weather conditions ($P < 0.0001$), but no interaction ($F < 1$).

scene became more foggy, subjects drove at faster speeds (Fig. 1b).

This finding suggests that the ‘blame’ for many such accidents may not lie solely in the irresponsible nature of the drivers but with an unfortunate quirk of our perceptual systems.

**Robert J. Snowden, Nicola Stimpson,
Roy A. Ruddle**

School of Psychology, University of Wales Cardiff,
Cardiff CF1 3YG, UK
email: snowden@cardiff.ac.uk

1. Sumner, R., Bagulay, C. & Burton, J. *Driving in Fog on the M4* (Department of the Environment/Department of Transport, TRRL, Crowthorne, Berkshire, UK, 1977).
2. Brown, I. *New Scientist* 24 December 1970, 544–545 (1970).
3. Thompson, P. *Vision Res.* 22, 377–380 (1982).
4. Blakemore, M. R. & Snowden, R. J. *Perception* 25, 34A (1996).
5. Thompson, P. in *Visual Motion and its Role in the Stabilization of Gaze* (eds Miles, F. A. & Wallman, J.) 29–52 (Elsevier, Amsterdam, 1993).
6. Distler, H. & Bulthoff, H. H. *Perception* 25 (suppl.), 58 (1996).

Further steps towards one culture

Consilience: The Unity of Knowledge

by E. O. Wilson

Knopf: 1998. Pp. 322. \$26

Paul H. Harvey

It was 1975. John Maynard Smith burst into my room. He had looked at E. O. Wilson's new book, the now infamous *Sociobiology*, and exclaimed that every Marxist in the Western world would be outraged by it. Maynard Smith was right, although the extent to which Wilson had anticipated the kerfuffle has never been clear.

Six major books on, Wilson brings us *Consilience*. Its first thesis is that the humanities, sociology, religion, ethics, art appreciation and almost everything else outside the current remit of science that has not found its analytical roots in evolutionary biology will soon do so. Its second thesis is that, once we understand how we came to be, we shall be constrained in determining where to go. Does this mean Wilson has failed to rehabilitate himself? The answer is no. The book is marked by a scientific rigour found wanting by many in his earlier works. Every step of the way he questions his assumptions, methods and tentative conclusions. Thankfully, he has not changed his vision, except by expanding its horizons.

Wilson makes a formidable argument for consilience — unifying on the basis of a common theoretical framework — of those near-disciplines that have no rational structure. The central concept underlying the book is the epigenetic rule: the idea that, because of our genetic make-up and the environment in which we develop, each individual has biases towards making certain behavioural decisions rather than others. If we can establish what those biases are and begin to understand their molecular and physiological causes, we shall be able to understand why we behave as we do. It is a massively tall order, and in terms of getting to grips with epigenetic rules, my own impression is that we have not advanced markedly in the past decade or so. Wilson still uses the examples of colour vision and incest avoidance, the empirical bulwarks of his co-authored book *Genes, Mind, and Culture* from 17 years ago.

But we have come a long way in understanding the functional analysis of behaviour, development and neurobiology. The book charts a marvellous path through the important empirical findings in these disciplines, placed in a framework for understanding and revealing tasks for the future that is a model of popular science writing. It also provides the strongest evidence yet for Wilson as a successful visionary. *Sociobiology* was an agenda as well as a review, and one



Reaching out: Wilson proposes that many disciplines will find analytical roots in evolutionary biology.

that has been largely realized. True, he got some parts wrong: his claim for the importance of trait group selection, for example. But most of it was right: evolutionary-based biological anthropology has emerged and even started to mature. He was not to know that molecular genetics would develop as it has, but its daily discoveries provide grist for his mill.

But in his new book, when discussing ethics, religion, sociology or art appreciation, he is forced to epigenetic rules viewed hazily, as through a fog (not one of his own making; that is the state of the art). The concepts are clear; if we could just summon the rules, we should reveal the proper foundations for the subjects. There is a long way to go, but we can afford to be patient because there is no competition. A proper understanding of these subjects will simply have to wait until we understand our own predispositions and the reasons for them.

In art, the few glimmers afforded by Piet Mondrian's spaced trees or the golden section point towards the rules but they cannot yet describe them, let alone trace their neural causations. (*Nature's* weekly art and science article is exactly what we need if we are to keep ourselves abreast of the challenge.) There are beautifully written patches of fluo-

rescent prose in Wilson's treatment of these subjects. He treasures apparent insights into Palaeolithic times and minds when the cultural changes occurred that produced us. He describes what he imagines are moments in the lives of our ancestors, often gleaned from days in the life of contemporary bushmen.

He is now the first to admit the 'just-so' nature of his tall tales, but we all have to start somewhere. It is getting a start that matters. There is no one way to proceed in science; one just has to go at it from whatever pinhole of opportunity presents itself, thinking around and around a subject until a way in is found. This view of the scientific enterprise is well known to evolutionary biologists and ecologists because they, like astronomers, often deal with the outcome of historical processes. They try to unravel history without an evident written record; in biology the fossil record can provide occasional guidelines for some problems, just as looking light years into the past can help astronomers.

The second theme of the book may still upset the die-hards. Wilson remains a determinist in an important sense. He believes that the epigenetic rules are sufficiently hardwired that our environment of choice is itself largely predetermined. We are content to behave only in certain ways and to live in

certain sorts of societies. A scientist wants to know why, but for the politician and the economist this limits courses of action. There will, of course, be alternatives available, but in Wilson's world they are appreciably more restricted than is generally thought. Economists do not appreciate the options because they have yet to think of them.

Wilson is a kind man, but he has no problems producing withering denunciations where they are called for. His description of our ecological plight, and his heartfelt plea for understanding and urgent action, are written with such authority, clarity and passion that I have read nothing to touch them. Nobel laureates in economics who fail to appreciate the world's finite resources and who argue for our wants in the short term rather than our needs for a sustained future are targets in this book. Evolutionary psychologists also get a timely warning. I was delighted to see one of their gurus pointing out so very dispassionately that, although Darwinian explanations abound, the theory is in its infancy and the database against which ideas are being tested remains poor.

Within this second theme, Wilson is at pains to circumscribe his own limits. Returning to art, he writes: "While biology has an important part to play in scholarly criticism, the creative arts themselves can never be locked in by this or any other discipline of science.... Works of art communicate feeling directly from mind to mind, with no intent to explain why the impact occurs. In this defining quality, the arts are the antithesis of science."

Wilson has had such a broad agenda through his career that he has frequently sought the advice of others for complementary expertise. For some of his research monographs he has collaborated with co-authors. At one end of the axis of excellence was Robert MacArthur. Fortunately, like George Oster, most of Wilson's collaborators lie close to MacArthur in this respect. And as with *Sociobiology*, so with *Consilience* — the vision is Wilson's but much of the detailed understanding of mechanism is necessarily gleaned from experts in their own fields. The responses from those working in the areas to which the clarion calls are directed may depend on how well Wilson has chosen his examples, and how accurately he has described the state of their art.

His own pool of knowledge and grasp of contemporary advances is astonishing, but even so we can expect the nitpickers to gather. I hope they will take time to evaluate the broader canvas and, where necessary, supply their own case studies in mounting whatever responses this monumental work demands. In any event, it is time to get on with the job. If we have to go through a bruising session like that following *Sociobiology*, I think that this time most evolutionary biologists will stand behind Wilson,

albeit obliquely. He deserves no less.

Some will argue that to bring evolutionary biology and the associated epigenetic rules into such areas as art appreciation will turn us into Philistines. They should be directed to Darwin's response to the evolutionary interpretation for life itself in the last paragraph of *Origin of Species*. The grandeur in his view of life provided us with a new dimension for appreciating the natural world but removed none of the old. Or, as Wilson might have it, our biophilia was intensified. □

Paul H. Harvey is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

At the limit

Impossibility

by John D. Barrow

Oxford University Press: 1998. Pp. 279.

£18.99, \$25

John L. Casti

To those enamoured of the belief that the human spirit knows no bounds or limitations, a tour of twentieth-century science must be a rather depressing experience. If there is any common denominator running through the scientific breakthroughs of our age it is the idea that there are limits.

Starting with Einstein's speed limit of the Universe in a ray of light and Heisenberg's limits on what can be measured with certainty, through Gödel's limits on what can be known by following a set of rules and on to Arrow's famous result about the impossibility of a perfect democratic society, the history of modern science is riddled with logical and physical limits of all sorts.

In this illuminating, well-written account of Limits (with a capital L), John D. Barrow chronicles and explains the limits of science as a reality-generation mechanism — and why it matters.

As he carefully points out, limits to the ability of science to answer a given question about the world as we see it come in all sizes and shapes. There are political and ethical limits on how much latitude a society is willing to give to investigators to answer a question. There are practical limits on the amount of time or energy or money that can be spent in answering a question. And, finally, there are logical limits.

Although Barrow pays more than lip service to the former sorts of limitations on our ability to tease out the 'scheme of things', it's clear that his real focus is on logical limits. Rightly so, too, because ultimately these are the limits that count — at least for the scientist and scholar, if not the politician and social agitator. When one is looking for a needle in a haystack, it's of more than passing interest to know that a

needle is really there to be found.

As Barrow points out, all the scientific knowledge we have about nature comes from models that we create of natural phenomena. These models, in turn, are almost always mathematical in character. It follows then that it's only a small approximation to say that the issue of logical limits to science comes down to the limits of computation, as any mathematical model can be regarded as an algorithm for processing inputs (the statement of the question and its circumstances) into an output (the answer).

This notion of scientific knowledge raises several fundamental questions. How does the mathematical model relate to the real-world phenomenon it purports to represent? Does the Turing-machine model of computation impose intrinsic limits on what we can know? What is the relationship between the computational powers of the human brain and those of our computing machines?

One question missing from the list — and the book — arises from the assumption that our models of nature must necessarily be mathematical. This question is: is there an alternative to a mathematical formulation of a model of natural phenomena?

The difficulty here is that one is left perched on the horns of a dilemma. Either you use the mathematical representation of the question of concern, and then try to justify why mathematical insolubility implies the same for the model's real-world correlate. Or you forsake mathematics altogether, and then face the difficulty of trying to create a convincing real-world substitute for the mathematical notion of proof, in order to produce a knockdown argument for why the question of concern is logically unanswerable. *Impossibility* evades this dilemma simply by ignoring the non-mathematical alternative.

But no matter. Taken on a whirlwind tour of the mind, Gödel's theorem, quantum theory, free will, voting paradoxes, time travel, computational intractability, sandpile models, the topology of space, nanotechnology, the forces of nature, the evolution of the Universe, complexity science, computer chess-playing, percolation theory, human consciousness, economic forecasting, black holes, the Brouwer fixed-point theorem, artificial intelligence and the Heisenberg uncertainty principle, one can only wonder how Barrow can possibly make all these things fit together into a coherent story about the limits to science. Well, contrary to all expectations, he does make them fit — and in only 250 pages!

So for about as good an account as you're going to get of where science stops, read this book. It won't tell you any final answer. But the journey is far more interesting — and important — than the destination. □

John L. Casti is at the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

Reinterpreting the historical record

Georges Cuvier, Fossil Bones, and Geological Catastrophes: New Translations and Interpretations of the Primary Texts

by Martin J. S. Rudwick

University of Chicago Press: 1997. Pp.301.

\$43.95, £27.95

Philippe Taquet

Georges Cuvier (1769–1832) was one of the outstanding figures in science during the first half of the nineteenth century. He managed to deal brilliantly with a threefold career in the areas of science, education and administration. Despite the convulsive movements of French history, he accumulated honours and distinctions, being praised to the skies by some of his fellows and hated by others.

Until recently, researchers and historians specializing in the sciences, particularly in France, have focused on two major aspects they considered typical of Cuvier's conservative thoughts, in order to criticize or mock them: his belief in the fixity of biological species and his adherence to the geological theory of catastrophism. Cuvier acted as a brake on the expansion of the transformist ideas developed by Jean-Baptiste de Lamarck, with his religious feelings influencing his interpretation of the history of the Earth.

Thankfully, over the past few years, some historians have offered a less limited view of Cuvier's thoughts. In 1964, William Coleman showed with great expertise, in a masterful book called *Georges Cuvier Zoologist*, that naturalists owe a tremendous amount to Cuvier for the monumental book he wrote on the comparative anatomy of the

animal kingdom. In 1995 Theodore Pietsch, the historian of ichthyology, published the first English translation of Cuvier's introductory text *L'Histoire Naturelle des Poissons* as part of his *Historical Portrait of the Progress of Ichthyology*. Pietsch rightly underlined the importance of Cuvier's work both in systematics and in the history of science.

Today, it is Martin J. S. Rudwick's turn to refute the erroneous or vague interpretations made of Cuvier's thoughts and to invite English readers to discover the most accessible of this French naturalist's main geological writings.

The text Rudwick introduces and comments on gives readers an opportunity to follow, step by step, the development of Cuvier's ideas, from his first, youthful writings to the last pieces he produced as a scientist who had become famous throughout Europe.

After providing a brief biography, Rudwick presents two little-known excerpts from letters written by the young Cuvier to his German friend C. H. Pfaff. They concern the geological observations Cuvier had made in Normandy and the reflections that had come to his mind after reading *The Theory of the Earth* by Jean-André Deluc. These texts are followed by "Mémorial on the species of elephants, both living and fossil" written in 1796; an article on the *Megatherium*, or "huge beast", from South America; a leaflet announcing in 1800 Cuvier's programme of research on fossil bones; a general summary and reconstruction of the skeletons of the different species of mammals from the gypsum around Paris; the famous report on the opossum from the gypsum beds in Montmartre; extracts of articles about living and fossil elephants written in 1806; a report on Noël André's *Theory of the Present Surface of the Earth*; the celebrated *Historical Report on the Progress of Geology since 1789, and on its*

Present State dedicated to the Emperor Napoléon; the first draft (1808) of the *Essay on the Mineral Geography of the Environs of Paris*; an article on "The fossil bones of ruminants found in superficial deposits"; the preface to the volumes on "Recherches sur les ossements fossiles", or "researches on fossil bones"; and the "preliminary discourse" which followed the preface, known as "The revolutions of the globe".

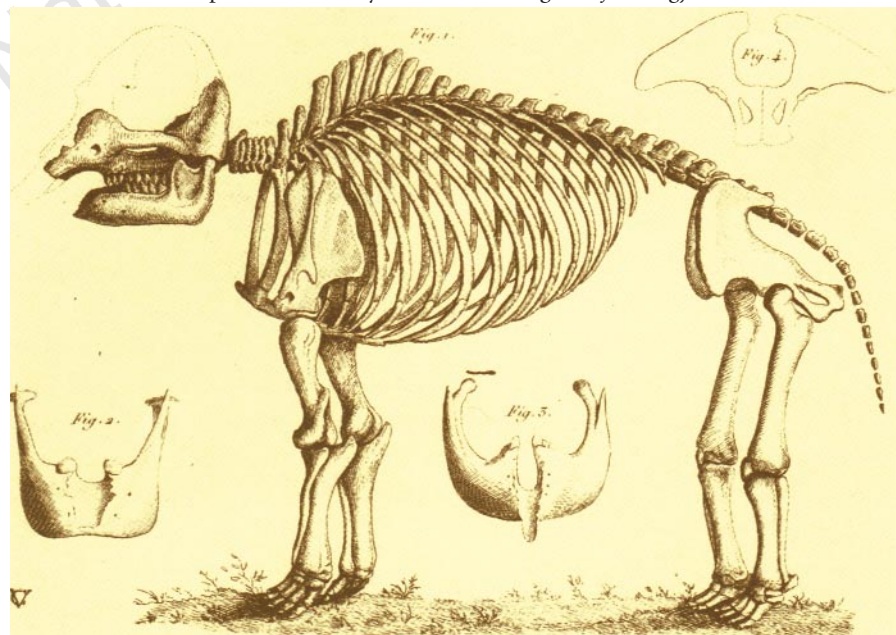
Finally, Rudwick offers two manuscripts by Cuvier: one from 1798 on "an animal whose bones can be found in the gypsum around Paris" (soon to be known as the *Paleotherium*), the other presenting Cuvier's notes for his lecture course in geology to the Lycée in 1805. Rudwick has translated and commented on a total of 19 important texts by Cuvier.

This book is very welcome because, since 1813, English-speaking historians have had at their disposal only Robert Jameson's translation of the "preliminary discourse". This translation is sometimes inaccurate, and, together with Jameson's comments, it has led, as Rudwick explains, "to the widespread belief that Cuvier had constructed his theories in order to support a literalistic interpretation of Genesis or to bolster the historicity of the biblical story of the Flood".

Rudwick's translations correct that gross misconception. But he hasn't just made a simple translation; he has examined each of Cuvier's manuscripts, outlining the corrections and modifications. As a specialist in the history of geology and palaeontology, he has made excellent comments on each of these texts, explaining the ideas and the methods developed by Cuvier, and giving further information about the scientific vocabulary Cuvier used. For instance, the word "analogue" (analogous) which Cuvier used was to be superseded by the word "homologous" introduced later by the British anatomist Richard Owen. The book also includes a bibliography of the sources used by Cuvier.

Rudwick's selection of texts is excellent, many of them, of course, being dedicated to fossil mammals. But I think two absent articles could have also been considered as crucial texts. The first is Cuvier's description of "the great fossil animal found in Maastricht quarries" which proves, for the first time and thanks to a perfect anatomical demonstration, the existence of an extinct species of a marine reptile, closely related to the varanid lizards today known as the *Mosasaur*. The second text is the one introducing the fossil skeleton of a flying reptile from south Germany, which Cuvier named "*Pterodactyle*". This was another decisive and revolutionary discovery, particularly in that it was made from the simple analysis of a plate published in a book, Cuvier himself having never seen the original piece.

The book will also, I hope, encourage everyone who can read French to immerse



Fossil bones revisited: Cuvier's reconstruction of a "mastodon", previously known as the "Ohio animal".

themselves in the original texts. Quite apart from his scientific skills, Cuvier was also a remarkable writer. Eager to convince, he had an accurate, clear, elegant style and a great sense of drama that allowed him to draw striking and even sometimes dramatic portraits of vanished worlds.

Throughout this book, Rudwick reveals the interest that can be found in reading the texts written by a man of science whose ideas and concepts deeply influenced nineteenth century society and who helped to raise the international status of anatomy and geology as professions. □

Philippe Taquet is in the Laboratoire de Paléontologie, Muséum National d'Histoire Naturelle, 8, rue de Buffon, 75005, Paris, France. e-mail: taquet@mnhn.fr

Radical remedies

The Antidepressant Era

by David Healey

Harvard University Press: 1998. Pp. 317.

\$39.95, £19.98

Leslie Iversen

Over the past 50 years the discovery of drugs that effectively treat the symptoms of schizophrenia, anxiety and depression has revolutionized treatment and led to radical changes in the way we view mental illnesses. David Healey's book focuses on the discovery and development of antidepressants and provides a fascinating insight into the history of this field. He skilfully interweaves his account of the roles played by the key scientists and clinicians with the powerful influence of pharmaceutical companies.

The first antidepressants were discovered, by accident, on each side of the Atlantic in the mid-1950s. The Swiss company CIBA tested imipramine in schizophrenia because the drug had been designed to resemble in chemical structure the first anti-schizophrenic agent, chlorpromazine. Although imipramine proved ineffective in the treatment of schizophrenia, an astute psychiatrist, Roland Kuhn, thought it had euphoriant or stimulant effects in some of his patients and he persuaded the reluctant company to test it in depression where it proved highly successful.

In the United States, meanwhile, anecdotal reports that the drug iproniazid had euphoriant effects in some of the tuberculosis patients for which it was developed were followed up by Nathan Kline, already famous for his work on reserpine in schizophrenia. Kline persuaded Roche to conduct a trial in withdrawn depressed patients which proved highly successful. Kline, a skilful publicist, released the story to *The New York Times* and lobbied senior executives at Roche for continuing support. Although iproniazid and other monoamine oxidase

inhibitors (MAOI) quickly fell out of favour because of their associated cardiovascular hazards (the 'wine-and-cheese' effect), they had a major impact as the first effective antidepressants in the United States. Kline was instrumental in lobbying Congress for funds to support psychopharmacology research and by the mid-1960s he had appeared on the cover of *Fortune* magazine and had become one of the ten best known men in America.

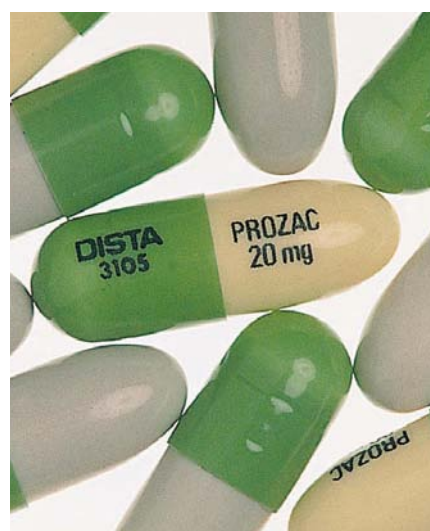
The MAOI were soon followed by amitriptyline, developed by the American company Merck and like imipramine because it chemically resembled chlorpromazine. The American psychiatrist Frank Ayd, however, had heard about Kuhn's results with imipramine and persuaded Merck to do trials in depression, which proved successful. Ayd also published an influential book on how to diagnose depression in general psychiatric practice, which helped to make amitriptyline the first antidepressant with major sales.

The drugs were expanding the diagnostic borders of depression, leading to the understanding that depression was far commoner than previously thought. This process reached its ultimate conclusion with Prozac and the concept that it could make some people feel "better than well".

Healey has perceptive insights into the role of the pharmaceutical industry in the development of the psychopharmacology era, and he portrays the often dramatic rises and falls in the fortunes of individual companies. He is at his best, though, in describing the shenanigans let loose by the drug era in his own profession, psychiatry. The advent of effective drug treatments led to major changes in this branch of medicine — introducing for the first time the concept of the controlled clinical trial and 'evidence-based medicine'.

There were often intense clashes between different ideologies (defined by Healey as "distinguished by their judgement of what ought to be regardless of what actually exists"). The psychoanalysts who dominated the profession in the 1950s and 1960s were pitted against the new biologically dominated school who embraced the use of the new drugs and called for radical changes in the classification of mental disorders. The two ideologies were often far apart — in many US psychiatric hospitals the psychotherapists refused to be associated with prescribing drugs for their patients, leaving this distasteful job to the "druggists" employed by hospitals. Eventually there emerged a backlash against the drug era in the form of "pharmacological Calvinism" — the belief that drug use is bad or even potentially dangerous if it makes you feel good.

The rise of psychopharmacology as a new field of research accompanied the drug era.



Superpill: Prozac-takers felt "better than well".

The "monoamine era" was heralded by the early discovery that reserpine acted by depleting monoamine neurotransmitters from the brain. This was followed by the finding that monoamine oxidase inhibitors and the tricyclic antidepressants (imipramine and amitriptyline) acted by making monoamines more available in the brain (by blocking the degrading enzyme or the tissue reuptake processes respectively). In the 1960s and 1970s the 'noradrenaline hypothesis' of depression dominated, although extensive attempts to provide evidence for the underactivity of brain noradrenaline systems predicted by this hypothesis in depressed patients all failed. Nevertheless, the hypothesis seemed to gain support from the discovery that new drugs that targeted noradrenaline reuptake (desipramine and maprotiline) proved clinically effective as antidepressants.

Twenty years later, however, this idea was conveniently forgotten as the selective serotonin reuptake inhibitors, of which fluoxetine (Prozac) is the best known, came to the fore. This apparent paradox perhaps reveals the naivety of the simplistic monoamine theories of the 1960s and 1970s, which reached their peak with the notion of one monoamine for each major illness: noradrenaline for depression, serotonin for anxiety, dopamine for schizophrenia and acetylcholine for dementia!

The antidepressant era represents one of the seminal events in the social and cultural history of the latter half of the twentieth century. This book is written in an individual and engaging style and the author reveals a deep knowledge of his subject; he has his own firm views but does not force them upon the reader. I found it a compelling read and hope that it will reach a wide audience. □

Leslie Iversen is in the Department of Pharmacology, University of Oxford, Oxford OX1 3QT, UK.

Objects of desire

The Elgin Marbles: Should they be Returned to Greece?

by Christopher Hitchens

Verso: 1998. Pp. 137. £11, \$17 (pbk)

The Elgin Affair: The Abduction of Antiquity's Greatest Treasures and the Passions it Aroused

by Theodore Vrettos

Secker and Warburg/Arcade: 1997. Pp. 238. £17.99, \$26.95

A. M. Snodgrass

The Elgin Marbles controversy, usually presented as a prize example of misplaced Greek emotionalism, in fact tells one far more about the British. The past two years have witnessed further instalments in this saga, which provide the occasion for a re-issue of Christopher Hitchens's short but powerful book of 1987, with the addition of a new foreword.

When it first appeared, a reasonable response to the book was that those who supported the return of the marbles to Greece had won the actual argument hands down. The heated and abusive language that they had once used against Lord Elgin had long since been usurped by their opponents. The retentionists, however, still seemed to hold all the high cards: possession of the objects, governmental and popular support, fear of awkward precedents and the serious constitutional obstacles to implementation.

A few of these elements have now changed. A Channel 4 (London) television broadcast by William G. Stewart in 1996 drew a startling response: among nearly 100,000 listeners who called in, more than 92% now favoured restitution. Work has begun on a splendid new museum in Athens where the marbles can be exhibited to full advantage. Above all, the British Labour Party — which, in opposition, had publicly supported restitution from the mid-1980s to the mid-1990s — came to power in May 1997. But word had earlier gone out that this former policy might be repudiated, and within hours it was. Is the debate therefore back at its starting point, when the first recorded proposal for the return of the marbles was put to the House of Commons in 1816?

Hitchens will have none of this. A high spot of the original book was (and remains) the shameful, ranting interview given by the then director of the British Museum in 1986, full of accusations of "cultural fascism". But the new foreword shows, from the most recent utterances of politicians in favour of retention, a return to an older vein which — in outraged propriety, historical inaccuracy and sheer complacency — has lost nothing



Frieze frame: the Elgin Marbles saga rolls on. Recent instalments include the start of work on a splendid Athens museum where the marbles could be exhibited to full advantage.

over nearly 200 years. For them, Elgin is still the altruistic saviour, the Greeks still unreliable denizens of an insecure third-world country. But Hitchens is no longer content merely to appeal to the historical record: instead, "morally certain" that the time will come when the marbles return home, he ends the foreword with an imaginary picture of that day.

For what readership, though, is Theodore Vrettos's book intended? It seems to belong to that late twentieth-century genre, 'faction' or 'docudrama', where extensive citation (indeed, at times pages of quotation) of contemporary documents alternates with mainly romantic speculations ("A glass of Portuguese brandy lay untouched on the small oak table ..."), often without a clear transition between the two — a kind of half-way house between Flashman and Antonia Fraser. All but the last few pages are set in the lifetimes of Lord and Lady Elgin. In so far as the book has a climax, it is not the removal of the marbles from the Acropolis (which is over before we are half-way through), but the trial of Robert Fergusson for adultery with Lady Elgin in 1808: one is left with the feeling that this is really the "affair" to which the book's title refers.

The research is haphazard — several of the extended quotations are misattributed — but it is rather the lack of background knowledge that is unsettling. No doubt there are readers who will be content to be told that Robert the Bruce was "the first king of Scotland", or that Sir Sidney Smith was besieging Napoleon in Acre rather than the other way round, but are they the kind of readers who will buy a book on this subject?

The brief concluding section is almost

entirely taken up with what is certainly the most interesting episode of the past 200 years, the brief flirtation of the London government, civil service and British Museum early in 1941 with the proposal to return the marbles to Greece once the Second World War was safely won.

Here Vrettos's account is fuller than Hitchens'. Alone in continental Europe, Greece still bore arms against the Axis powers, and the offer to return the Elgin Marbles was seen as the right gesture for its British allies to encourage it to maintain that heroic stand. Remarkably, the Foreign Office seems to have been the moving spirit behind the idea, and for a moment it looked as if a way could be found around the legal, practical and scholarly obstructions to such an action. (Moral arguments played no part: Lord Elgin's action was still held to have been "right in every way".)

Although Vrettos is on record as having taken up a position on the larger question elsewhere, he avoids doing so here (except perhaps in the emotive phrasing of his subtitle). The main service of his book is to bring out fully the profound sadness of the whole Elgin episode. Here was a typical aristocrat of his time, whose appointment to the Embassy at Constantinople, at a uniquely favourable political juncture, had given him the opportunity of greatness. Instead, he had to watch the disappearance in succession of his nose (eaten away by an infection), of his fortune, of his career prospects (through his internment by the French), of his wife's affection and, above all, of his reputation in perpetuity.

It was his misfortune that an act of destructive appropriation, which otherwise

might have been remembered only by scholars, attracted uniquely widespread condemnation, with Byron's voice not merely soaring above the contemporary chorus in Britain, but speaking to Greeks as well, and to every later generation in both countries.

The kind of internationalist spirit that protested then is still quietly at work today: it will never go away. In the long run, there can only be one end to such a story. □

A. M. Snodgrass is at the Faculty of Classics, University of Cambridge, Cambridge CB3 9DA, UK.

Civilized living

Cities for a Small Planet

by Richard Rogers

Faber: 1997. Pp. 180. £9.99 (pbk)

Roy Porter

Cities are certainly the most conspicuous symptoms of our global environmental crisis, and perhaps even the chief cause. One statistic will here suffice — with its 20 million inhabitants and four million cars, Mexico City is growing at the rate of 80,000 people a month — and also one grim reminder: all previous great urban civilizations, from those of the Indus valley, the Tigris and Euphrates onwards, have utterly collapsed.

Cities for a Small Planet, an exquisitely designed and illustrated reworking of Richard Rogers' 1995 BBC radio Reith lectures, tackles the environmental costs of cities, while affirming their human potential. Cities are meant to bring civilization, but all too often today it is pollution, poverty, crime, congestion and chaos that they call to mind: hence the flight to greenfield sites. So: what future for our cities? And specifically, asks Rogers, where do architects stand in all this — are they sinners, scapegoats or saviours?

Recent decades have seen cities compromised by market-driven developments — sterile zoning into 'single-minded' spaces such as business parks and out-of-town shopping malls — and wounded by the great gashes made by highways. Today's megacity may be bejewelled by the odd breathtaking postmodernist tower, but (especially in the United States) it has turned its back on the public areas people crave and urban vitality demands. All too often, Rogers admits, the architect or planner has been the villain of the piece, or at least supinely complicit in the making of these urban wastelands.

How could we expect otherwise, when the profession is typically paid to produce not integrated spaces but blocks, in one-off commissions from property developers? "Buildings of all types are packaged and standardized," Rogers concedes. "Architects are selected for their low fees rather than for the



Under a cloud: are sprawls such as pollution-plagued Mexico City heading for disaster? Four million cars congest its streets and its population of 20 million is growing at the rate of 80,000 people a month.

quality of their work. The profession is condemned to turning out the largest enclosure for the least money in the shortest time ... these buildings are the energy-guzzling structures that are consuming half of the world's annual energy." The result? Buildings that enrich clients but impoverish the quality of urban living.

In place of today's characteristic *laissez-faire* sprawl, Rogers argues the case for the compact city, one with multiple and overlapping functions, and endowed with ample people-friendly public sites. To achieve this, public will is needed, as in the showcase post-Franco renewal of Barcelona in the 1980s under the populist mayor Pascal Maragall. By contrast, London, rudderless since the abolition of the Greater London Council, is collapsing amid worsening traffic congestion and pollution which carry economic as well as health costs.

Architect-planners have their role to play too. Given today's desiderata of sustainable growth and environmental renewal, the city and its structures must be rethought on viable ecological principles. Essential to that are energy-efficient buildings, whose specialized technical features will use natural resources (trees, sunlight, wind) and minimize waste.

Escorting the reader through a portfolio of buildings and developments mainly designed by his own partners, Rogers shows what has been done (efficient use of the atrium, for instance) and what remains to be achieved. It is disappointing that his most environmentally sympathetic proposals — such as those for Shanghai — seem to be the ones that never got off the drawing-board,

presumably because they contravened the developers' demand for profit-maximization and the conventional urban-political axiom that the motorist always comes first.

Londoners in particular will welcome another chance to ponder his imaginative plans for Trafalgar Square (now, as he remarks, just a vast roundabout) — though they may not be able to avoid a wry smile, given his own involvement with the Greenwich Dome, when Rogers says that, by way of a Millennium celebration, a "series of local projects" would be far preferable to a "sweeping Beaux Arts masterplan".

Mammon of course has always been the ultimate foe, but the most destructive force this century has been the car. Every redesigning of the city to make it subservient to the needs of private vehicles has been another nail in its coffin as an agent of human congregation: all too often, all that is left of public space is the unsafe, frenetic pavement. Rogers is at his most passionate in crusading to reclaim the city from the car — whether through big improvements in public transportation systems or, better still, by cutting the need for journeys themselves, for example by 'living above the shop'.

Read this book while stuck in the traffic jam on the way to work. Think of all the friendlier urban environments in which you might be reading it. And then make waves towards putting into effect the ideas so imaginatively, so ardently advanced. We owe it to the next generation that it should have no mean city to inherit. □

Roy Porter is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

Cyber-sociology

The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration

by Robert Axelrod

Princeton University Press: 1997. Pp. 232.

\$49.50, £35 (hbk), \$18.95, £14.95 (pbk)

Karl Sigmund and Martin A. Nowak

When Robert Axelrod's book *The Evolution of Cooperation* appeared in 1984, it turned into an instant classic, and deservedly so. Its influence reached far beyond political science, experimental psychology and evolutionary biology; by now, a large public has become familiar with its clear and soberly optimistic message. Richard Dawkins has hailed it as "one of the two books that have excited me most" (the other was one of his own).

After such a triumph, what do you do for an encore? It must have been a daunting task to write a sequel. Axelrod decided upon a totally different format. *The Evolution of Cooperation* was tightly focused on one model (a population of individuals interacting in repeated 'prisoner's dilemma' games) and explored one issue only, namely reciprocal aid. *The Complexity of Cooperation*, on the other hand, is a loose collection of a handful of papers published in diverse journals and dealing with sundry aspects of exploring new strategies, converging on norms, building coalitions or disseminating cultural traits. Each paper is preceded by a short introduction.

The common thread running through all these chapters is the so-called 'bottom-up approach', which is by now quite orthodox: it consists in reducing the interaction to a simple game, devising programs for playing it, and running computer simulations of populations of agents guided by these programs.

Axelrod explains that the "complexity" in the title has a double meaning: the interactions that he examines are complicated, and the techniques he uses are those of complex adaptive systems theory in the sense propagated by the Santa Fe Institute and its aficionados.

Despite the title, one will not find lots of complexity in this book. The hype surrounding the 'edge of chaos' and 'self-organized criticality' is totally eschewed, and the simplicity of the models is occasionally breathtaking.

Axelrod starts on his home turf, with the prisoner's dilemma game, where each player does better by choosing to defect, no matter what the co-player is doing, with the result that mutual defection (rather than the more rewarding mutual cooperation) gets established. The author's celebrated computer tournaments for the many-rounds game, where cooperation won, were based on a

narrow sample of some dozen strategies submitted by eminent experts. But Axelrod later adapted the genetic algorithms of his colleague John Holland to create and test a large set of new, randomly generated strategies.

This ground-breaking work from the mid-1980s, which is reprinted as the first chapter in the book, remains one of the most elegant and convincing examples of genetic programming. Somewhat disappointingly, Axelrod does not elaborate in his introduction on the remarkable subsequent development in the field of reciprocal aid. Interested readers will have to refer to other publications instead — Axelrod's own periodical reviews, for instance, are considerably better documented.

The same criticism applies to the chapter on promoting norms (including 'meta-norms' demanding to punish those who disobey norms): Axelrod does no justice to a wealth of later developments (by Boyd, Sugden and Young, to name but a few) which were stimulated to a large extent by his own work. Instead, the reader is offered detailed information on the various political committees on security, arms control and so on, using Axelrod's expertise: a mild form of the syndrome plaguing many political scientists since Henry Kissinger's heyday.

It is on this question of practical impact that he becomes less convincing. By reading Axelrod, politicians can obtain (like everyone else) a better understanding of the game they are playing, but they will not become better at playing it. Axelrod's abstractions should be used as thought experiments only, not as flight trainers.

Consider, for instance, Axelrod's spin-glass model for choosing sides in a political conflict. It is essentially a physicist's worldview: the nation states have a certain propensity to align with each other on the basis, say, of ethnic, religious, territorial, governmental and historical issues. The nations can change side one at a time, thereby reducing their frustration. One can compute which alignments cause minimal frustration in the whole system.

Axelrod does this for the Europe of 1936, and finds two stable configurations. One consists essentially of the Soviet Union against the rest of Europe, the other of the Axis powers against all comers. A couple of countries misbehave but, for a 'prediction' of the actual alliance, this is pretty good. As Niels Bohr used to say, however, it's "predicting in advance" where things become hard. Besides, Winston Churchill offered an even simpler explanation of Europe's politics which carries greater conviction: in his view, the Second World War was just the continuation of the First World War, with a 20-year truce in between.

This is not to deny that Axelrod's spin-glass diplomacy is ingenious, and helps in approaching 'what if' questions (for

example, what if some territorial dispute had been settled, or some country not remained neutral). In fact, it should serve as a gold mine for political scientists.

The same holds for Axelrod's tribute model, which displays the emergence of major powers and their dissolution by imperial overstretch, as well as for his lattice model on the dissemination of cultural traits, which exhibits a fascinating and thoroughly eye-opening interplay between local convergence and global polarization. In each case, Axelrod manages to find a minimalistic model with a maximum of interesting features. He can afford blissfully to neglect previous work (for instance, on the game theory of coalition formation, or on rational behaviour) because his original approach is often more to the point. Each of his models is a first step in a promising direction.

The knack for simplicity seems almost an instinct with him; this instinct also tells him to stop before complexity really sets in. □

Karl Sigmund is at the Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Vienna, Austria. Martin A. Nowak is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

Bar fun

Ripples on a Cosmic Sea: The Search for Gravitational Waves

by David Blair and George McNamara

Allen and Unwin/Addison-Wesley: 1998.

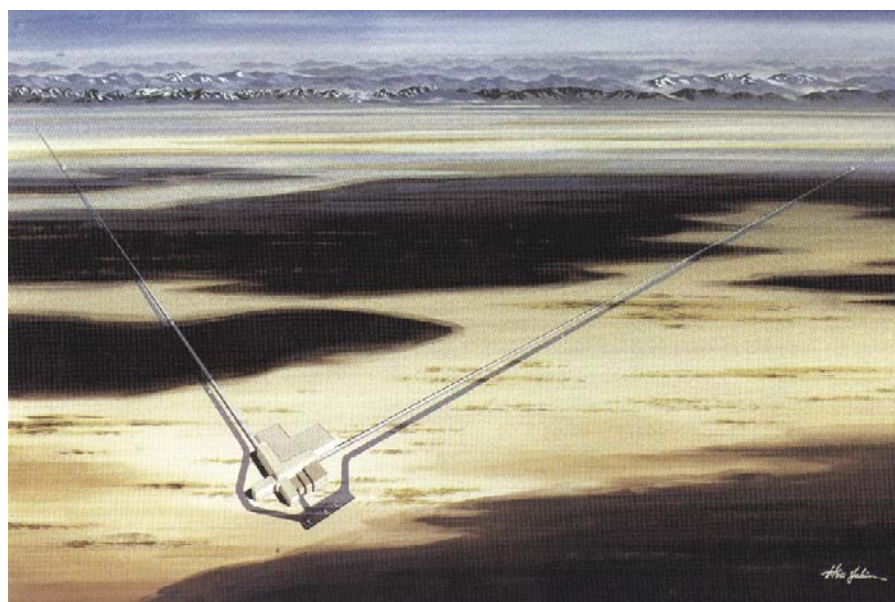
Pp. 179. Aus\$17.95, £7.99, \$22 (pbk)

Harry Collins

Spin a nuclear submarine about its short axis until it is near to breaking and the general theory of relativity says it will emit gravitational radiation of about 10^{-24} watts. This is not much; an ant walking up a wall uses 10^{-7} watts.

The comparison is found in *Ripples on a Cosmic Sea*, and shows why, to outsiders, the search for gravitational waves seems almost crazy. Nevertheless, the US National Science Foundation has funded a joint project between the California Institute of Technology and the Massachusetts Institute of Technology that uses two huge interferometers to look for such shivers in space-time. The interferometers are like Michelson-Morley experiments but with arms 4 kilometres long. Similar but smaller devices are being built in Europe and Japan, and David Blair's group is part of an Australian collaboration that has made a start on another, near Perth.

The interferometers are only the latest stage in a story that started in the 1960s when Joseph Weber of the University of Maryland put together the first resonant detector. He hung bars of aluminium weighing a couple of tons inside vacuum chambers, insulating them from all known influences. When two



Ripple detector: artist's impression of the Laser Interferometer Gravitational-Wave Observatory, taken from the LIGO website (<http://www.ligo.caltech.edu>).

bars separated by thousands of miles 'rang' in coincidence, Weber argued that gravitational waves were a putative cause of the disturbance; the source could be cosmic catastrophes such as supernovae.

By the early 1970s, Weber's claims had become so forceful that others built similar devices, but by the mid-1970s most people thought Weber was mistaken. Blair and George McNamara provide a colourful description of a confrontation between Weber and Richard Garwin at a conference in the 1970s that might have come to blows had the chairman not stepped in. Weber continues to press his claims, but most of the rest of the field moved on long ago.

The next generation of detectors were Weber bars cooled to liquid-helium temperatures and below. Blair himself runs such a device, although his bar is niobium rather than aluminium. Such attempts to detect gravitational waves are dealt with in the last third of the book. The first section is a short history of science, taking us through Newton to general relativity and the curvature of space, and in the middle is a section about potential sources of gravitational waves. The 'nuclear submarine problem' makes it impossible to generate detectable fluxes on Earth, so we must look to cosmic sources. The huge curvatures and energies associated with cavorting neutron stars and black holes make them the current favourites.

Astronomers Joseph Taylor and Russell Hulse studied the decay of a pair of orbiting pulsars over 20 years. The slowdown of 70 microseconds per year in an orbit of seven-and-three-quarter hours fits well with the predicted loss of energy through gravitational radiation. The 1993 Nobel prize for physics was awarded to Taylor and Hulse for this first indirect confirmation of the existence of the waves, but direct observation is

still the holy grail.

Towards the end of their orbital decay (in about 300 million years for the Taylor–Hulse pair), neutron stars nearly touch and circle hundreds of times per second. The final inspiral should produce a characteristic 'chirrup' of gravitational radiation of enormous power, about as bright as 100,000 galaxies. But even these immense fluxes will bend space so little that they will be on the edge of detectability by the most advanced interferometers now being designed. Colliding black holes should be even more fun, and there ought also to be a just-detectable background of gravitational radiation left over from the Big Bang.

Ripples on a Cosmic Sea is currently without competitors. However, parts of the book are so relentlessly 'popular' as to be patronizing. The most informative chapters are those dealing with pulsar sources, and those describing detectors and their inventors. Chapter 12, on laser interferometry, is especially good. It presents important ideas clearly, not shirking complicated new techniques. It also mentions the more operative organizational upheavals that have attended the US programme and the dirty dealings associated with the funding of different national projects. These chapters provide a nice account of sources and technology and a good thumbnail sketch of the field's history. The book could then be passed on to a young niece or nephew.

Those who get a taste for the science of gravitational radiation detection from the book might want to borrow a library copy of Peter Saulson's *Principles of Interferometric Gravitational Radiation Detectors*, which includes a short section on resonant bars, or Blair's edited collection of technical essays *The Detection of Gravitational Waves*. In both books the equations are supplemented with

good clear writing, but you'll need to be rich to buy them. □

Harry Collins is at the Centre for the Study of Knowledge, Expertise and Science, University of Wales at Cardiff, 50 Park Place, Cardiff CF1 3AT, UK.

e-mail: collinshm@cardiff.ac.uk

Disregarding the social sciences

Is the Temperature Rising? The Uncertain Science of Global Warming

by S. George Philander

Princeton University Press: 1998. Pp. 258.

\$35, £22.50

Hans von Storch

'Global warming' has become a household term; it needs no explanation when used in the news. 'Is the temperature rising?' is a question of utmost interest to anyone concerned about the global environment. Both phrases announce the subject of this book.

The volume is based on a course for undergraduate students given by the author at Princeton University, so it is also well suited to educated laypeople. It is written clearly and contains informative figures. The extensive appendices provide additional technical material, and an eight-page glossary is included.

The author explains complex scientific concepts in a precise language and with delightful illustrations. Of the waves on the ocean's surface, he writes: "A breeze that blows over the ocean, like a bow that strokes a violin string, readily excites music in the form of waves. The audible sound of a violin is soothing when the pressure from the bow is gentle and becomes strident when the pressure is great. The ocean's music changes similarly from ripples ... to foaming, lashing waves as the gentle breeze grows stiff ... and becomes a gale that whips the ocean into frenzy." He describes the weather as the "music of our sphere" and explains the principle of chaos using a thought-experiment of a skier losing his or her wallet on a slope. The book is a pleasure to read.

But its title is deceptive. Most of the volume explains the workings of the climate machinery and its components such as radiation, clouds, weather and oceans. The question 'Is the temperature rising?' is dealt with only on two pages in the last chapter.

The answer is 'yes', which will be no surprise to anybody who knows the observational record compiled by the University of East Anglia in England. The more interesting question 'Will the temperature continue to rise?' is dealt with in even fewer lines: the author quotes "very probable" rates of increase of 0.5–2 °C until about 2050, from

studies of climate models forced with continually increasing greenhouse gas concentrations.

The author reduces the uncertainty of climate science mostly to the fundamental problem of nonlinear dynamics and the related problem of determining accurate initial conditions. Other fundamental aspects are neglected, such as the climate system's phase space with infinitely many degrees of freedom, the openness of the system and the need to impose various semi-empirical parameters on climate models. As a result, the book comes across as over-optimistic about the task of analysing and modelling the climate system and its sensitivity.

As to the science of global warming, the author takes this to mean the purely physical sciences, exploring myriad processes that take place in the atmosphere, the oceans, the lithosphere and so on. Thus a planetary-scale view of global warming emerges, with relocated deserts, rising sea level, a defunct Gulf Stream and more El Niño events. What these changes mean for humankind is not specified. Instead, the author stays vague, mentioning the "considerable inconvenience over the next few decades for our particular species" and issuing general warnings about forthcoming unspecified "dangers".

The science of global warming should, however, also deal with the specifics of these dangers: that is, it should look at the possibility of abating and adapting to them. And as ecosystems and people experience global warming mostly locally, this analysis must be done on the local scale.

The author mentions Thomas Malthus's prediction, 200 years ago, that England was heading towards "gigantic inevitable famine" because of the steady increase in population. Perhaps we should learn from Malthus's science that the fate of the environment and humankind is only partly determined by natural dynamics, such as population growth and the carbon cycle, and that social dynamics and technological innovations are important as well.

Every scenario of possible future climate change is based on scenarios about the anthropogenic output of greenhouse gases: that is, on the outcome of a highly complex social system. The challenge of climate science therefore is to integrate the social sciences — by coupling economics models, diagnosing the socio-scientific construction of the 'climate problem' or analysing the impact of climate policy on people's well-being. Unfortunately, the author has chosen not to take up this challenge; and where he refers to old misconceptions, such as the infamous "climate and civilisation" hypothesis of Huntington, he fails to put them in proper perspective. □

Hans von Storch is at the Institute of Hydrophysics, GKSS Research Center, PO Box, 21502 Geesthacht, Germany.

Thumbs up for signal work

The Expression of the Emotions in Man and Animals

by Charles Darwin

Third edition, with introduction, afterword and commentaries by Paul Ekman.
HarperCollins/Oxford University Press: 1998.
Pp. 473. £16.99, \$30

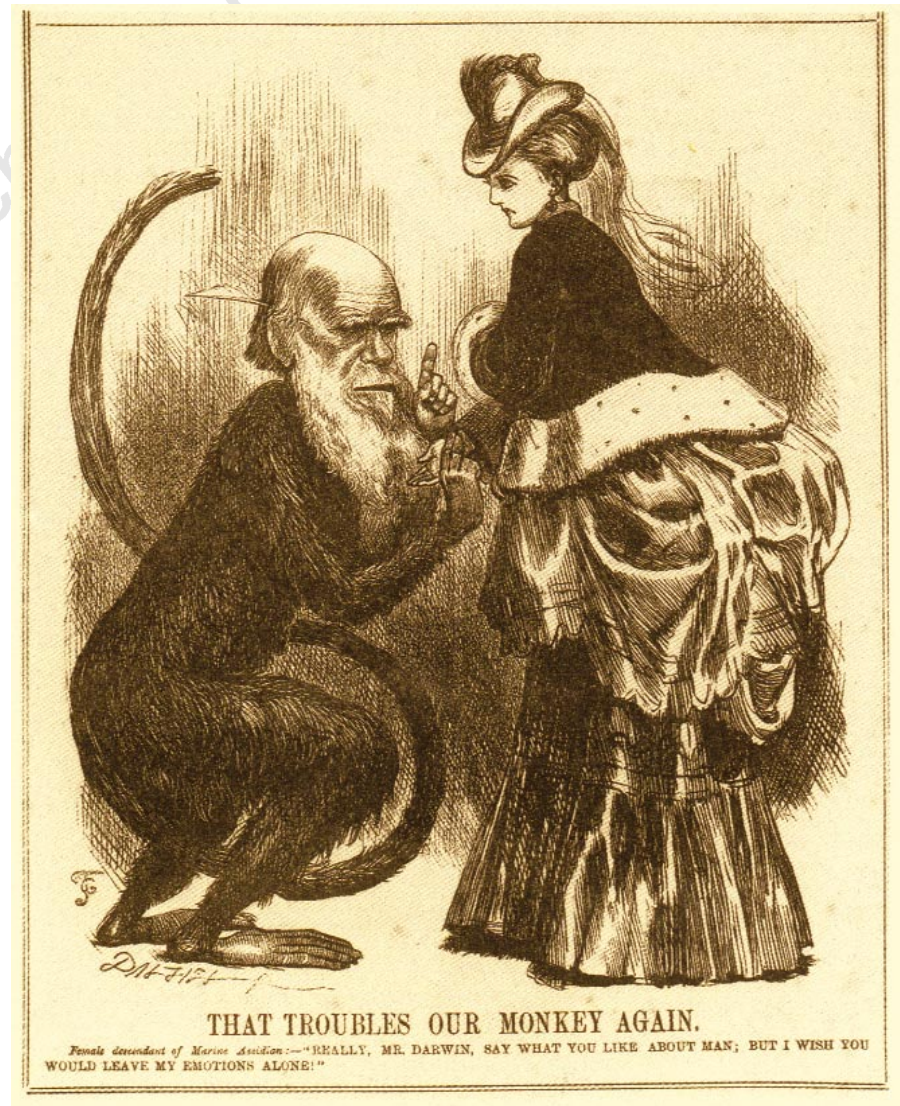
Simon Baron-Cohen

Almost everyone has heard of Darwin's *The Origin of Species*, and many will have heard of his book *The Descent of Man*. These books disseminated the theory that now stands as the framework for the biological sciences. But his book *The Expression of the Emotions in Man and Animals* is less well known, except possibly among readers in the field of psychology. This book is where Darwin laid out his highly original thesis that emotional expressions in humans have their counterparts in the behaviour of other species, suggesting an evolved commonality of form and function.

The first edition came out in 1872 and sold 9,000 copies in the first four months. The text contains a wealth of detail from a wide range of primate species, describing the expression of fear, anger, contentment and sadness. It relates these to observations of the behaviour of human infants. Some of the observations were drawn directly from his descriptions of his own child: for example, where he details the expression of guilt on his son's face when the boy was just two years old.

The second edition of the book was edited by Darwin's son Francis, and appeared in 1889, seven years after Charles Darwin died. Francis appears to have brought it out because Charles had prepared many revisions and additional material not available in the first edition.

Now, more than a century later, the world's leading Darwin scholar on the subject of the emotions, Paul Ekman, has produced a treasure trove of a third edition. Ekman is widely known in this field for his empirical contribution to defining the musculature of facial expressions of emotion, effectively rendering what was once a phe-



Seconding that emotion: she may not have liked it, but Ekman's findings back up Darwin's work.

nomenon characterized only by subjective description (and thus difficult to study scientifically) into a branch of physiology where objective measurement could be practised.

Ekman's own work has not been confined simply to validating the muscular changes in emotion, although this has been a major advance in the field. He has also studied both recognition and expression of emotion in every aspect, but perhaps most famously in cross-cultural studies. His finding that a core set of emotions are universal, irrespective of culture, is of course a result perfectly predicted from the Darwinian theory.

It might amuse readers to hear that Ekman set out to study emotional recognition in far-away tribes in non-literate parts of the world just to prove Darwin wrong. As the results turned out, Ekman's *a priori* cultural determinist view was refuted, Darwin's biological determinist view was upheld, and the results steered Ekman into a career that generated an enormous wealth of new data in this field.

Ekman has produced this third edition so as to bring Darwin's book into contact with this new data. Darwin's ideas were mostly based on anecdote, correspondence, some observation and speculation. Many of these ideas have been confirmed by later work, and Ekman's editorial commentary throughout the text helps the reader to sort out which of these have stood the test of time, and which Darwin simply got wrong. To ensure they are balanced, commentaries are also based on correspondence from a distinguished panel of international researchers working on emotion.

This new text is scholarship at its best. The strength of Darwin's writing still shines through, as well as his drive to explain the form of each emotional expression: why is blushing associated with embarrassment? Why do we purse our lips when we concentrate? But Darwin's own account is now properly set in the contemporary scientific context. The commentaries appear in boxes within the text, producing the effect of a dialogue between Darwin and modern science, bridging the century.

The publishers immodestly include a subtitle on the front cover that reads "Definitive Edition". This is a fair description as we leave the twentieth century, although it might be seen as a touch myopic by reviewers of future editions. Time will tell. But without doubt, this new book will be required reading for Darwin scholars of emotion. □

Simon Baron-Cohen is in the Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK.

Related book

A new paperback edition of Charles Lyell's *Principles of Geology* has just been published (Penguin, £9.99). Darwin constantly

referred to this work while voyaging on *The Beagle*. It has now been edited into a single volume by the historian of science James Secord, who also provides an introduction outlining Lyell's writing strategy and an explanation of the book's enormous cultural impact. A bibliography of reviews is also included.

A feast for the senses

Tibaldo and the Hole in the Calendar

by Abner Shimony

Copernicus: 1998. Pp. 165. \$21, £15.50

Stephen Battersby

Tibaldo Bondi lives in Bologna, Italy, in the late sixteenth century. In 1582 (the year is significant), we find him at the age of 11 studying at the school of St-Joseph-in-the-Corner, hoping one day to become a great doctor. He is very intelligent and very sensible; a little too good, perhaps. But one day in a Latin dictation he translates a proclamation from the Pope that horrifies him: Gregory XIII has decided to replace the Julian calendar, whose imprecision means that calendar holidays are gradually slipping from their proper season, with a better one concocted by a commission of astronomers.

This is sensible, and satisfies Tibaldo. But, to make up the difference in time accumulated over one-and-a-half millennia, there will be no 15–25 October that year — there will be a hole in the calendar, and Tibaldo's twelfth birthday will fall through it. He becomes obsessed, and sets out with great resourcefulness to fight the calendar and get his birthday back.

This is a lovely book. It is an Improving Tale for children, but not excessively moralistic; it is written in an austere style, but one that allows for some dry humour. And it is beautifully produced, with a cover and illustrations drawn in stylized Renaissance cartoon fashion by the author's son, Jonathan Shimony. They, too, are often playful, and worth lingering over.

On the whole it is also an adult book — Abner Shimony writes about the power of propaganda and the details of childbirth, and he doesn't avoid using difficult words where they are needed. The aim is unashamedly to teach, but the story allows the author to do this naturally: the problem of constructing a consistent calendar leads us to astronomy, especially the astronomy of the late Renaissance. Tibaldo's studies lead us to the medical science of the time; and his quest takes us into the politics.

In every case, we are shown the virtues of common sense. At a time when the natural philosophers of Europe were beginning to appreciate the power of experiment and



observation, many of Shimony's ordinary people already have a practical sense that can almost be equated with empirical science.

We see this in the early description of Tibaldo's father, Lorenzo, at work. Once the distinguished professor of medicine, Turisanus, has given his pompous and meaningless diagnosis (which Shimony describes ironically as "the most important part of the treatment"), Lorenzo sets about helping the patient. He cleans and dresses wounds, or sets bones — the things that he and his predecessors have learned from experience. More explicitly, Tibaldo's older sister Anna Maria, a midwife, advises him "to observe which treatments work, and which do not". Tibaldo is very lucky to be surrounded by such a sensible family, disdainful of superstition, and benefiting from a few centuries of authorial hindsight. Yet somehow the result isn't implausible.

But what about the story? Does Tibaldo succeed in his quest, repair the hole in the calendar and enjoy his birthday feast after all? Or must he grow up and abandon his obsession? I'm not telling. Read it. □

Stephen Battersby is an assistant editor at Nature.



The core–mantle boundary layer and deep Earth dynamics

Thorne Lay, Quentin Williams & Edward J. Garnero

Recent seismological work has revealed new structures in the boundary layer between the Earth's core and mantle that are altering and expanding perspectives of the role this region plays in both core and mantle dynamics. Clear challenges for future research in seismological, experimental, theoretical and computational geophysics have emerged, holding the key to understanding both this dynamic system and geological phenomena observed at the Earth's surface.

The Earth acquired its primary layered structure—consisting of a molten metallic alloy core overlain by a thick shell of silicates and oxides—very early in its history, and the region near the core–mantle boundary (CMB) surface has undoubtedly played a significant role in both the core and mantle dynamic systems throughout their subsequent 4.5-Gyr evolution. The CMB has a density contrast exceeding that found at the surface of the Earth, has contrasts in viscosity and physical state comparable to those at the ocean floor, and has much hotter ambient temperatures than found near the surface^{1,2}. These factors, together with the requirement that significant heat be flowing from the core into the mantle in order to sustain the geodynamo (the core magnetohydrodynamic flow regime that produces the Earth's magnetic field), provide the basis for the conventional view that a significant thermal boundary layer exists at the base of the mantle, with a temperature contrast of $1,000 \pm 500$ K (refs 2–4). With ambient temperatures averaging around 3,000 K, this hot thermal boundary layer is a likely source of boundary-layer instabilities, and it has been speculated that upwelling thermal plumes ascend from the CMB to produce surface hotspot volcanism such as at Hawaii and Iceland, transporting about 10–15% of the surface heat flux^{5–7}. The large CMB density contrast favours a mantle-side accumulation of chemical heterogeneities derived from initial and continuing chemical differentiation of the mantle, possibly including downwelling subducted slabs of former oceanic lithosphere; distinctive chemical signatures of these heterogeneities may eventually be entrained into thermal plumes, resulting in unique hotspot chemistry^{8,9}. There may also be a core-side chemical and thermal boundary layer, but the low seismic velocities and molten state of the core make it much harder to analyse any structure in the outermost core².

In the past decade, seismic tomography has provided steadily improving images of deep mantle elastic velocity heterogeneity, revealing the presence of significant large-scale ($>2,000$ km) patterns of coherent high- and low-velocity regions at all depths^{10–12}, with enhanced lateral variations in the lowermost 300 km. Attributing the seismological variations to lateral thermal and chemical heterogeneity in the boundary layer suggests several mechanisms of coupling between the core and mantle, influencing the core flow regime as well as irregularities in rotation of the planet^{2,13–15}. An experimental demonstration that chemical reactions may be continuing between the core and mantle has raised the possibility of electromagnetic coupling across the boundary^{16,17}. These considerations assign possible importance to the CMB layer for dynamic processes throughout the planet (including an intimate relationship to surface structures), yet the precise role of the CMB region remains a topic of vigorous research and debate requiring improved resolution of the structure and processes operating there.

Several recent seismological discoveries have revealed new attributes of the mantle-side boundary layer overlying the CMB. The lowermost 200 km of the mantle (known as the D'' region) has

long been characterized by seismologists as distinctive in its properties from the overlying deep mantle. Numerous studies have revealed the presence of intermittent stratification of the D'' region in many areas, with 1.5–3% velocity discontinuities for both compressional (P) and shear (S) waves at depths of 150–300 km above the CMB^{18–21}; studies have also shown the presence of laterally varying shear-wave anisotropy (birefringence that splits the S waves into orthogonal polarizations that travel with different velocities owing to organized mineralogical or petrographical fabrics) in the D'' region that is, at least sometimes, related to the discontinuity structure^{22–26}. Both observations lack satisfactory explanations in terms of current understanding of thermal and

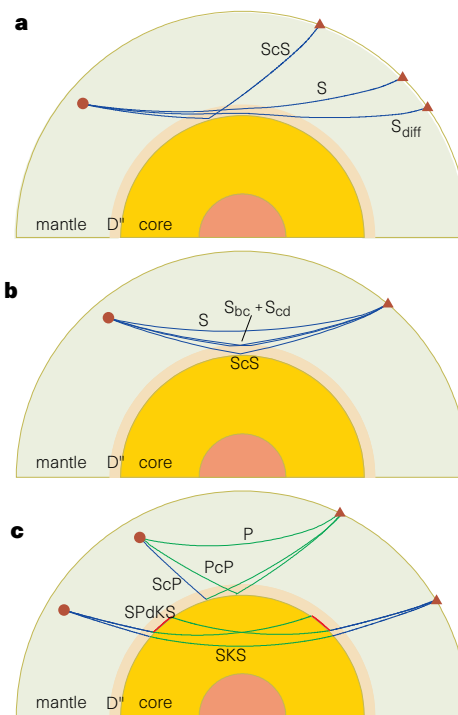


Figure 1 Seismic-wave ray-paths from a deep focus source (circle) to a receiver (triangle) for phases that have been extensively used to image the velocity structure of the deep mantle. **a**, Core-grazing phases such as the reflection ScS, direct S near 90°, and diffracted S (S_{diff}) in the shadow zone. These have P-wave counterparts. Grazing waves provide sensitivity to lateral heterogeneity and anisotropy in the D'' region. **b**, D'' triplication phases, including reflections (S_{bc}) and refractions (S_{cd}) from a velocity discontinuity. These have P-wave counterparts. Triplications constrain the depth and strength of any discontinuities. **c**, Phases used to study the ULVZ include short-period reflections (PcP) and conversions (ScP), as well as the long-period SKS phase and its associated SPdKS phase (involving P energy diffracted along the core).

mineralogical structure of the deep mantle. A newly discovered laterally varying ultra-low-velocity layer in the lowermost few tens of kilometres of the mantle^{27–29} is also challenging the conventional view, as the magnitude of the velocity drop (10% or more reductions in both P and S velocities) is so large that it probably involves partial melting of the mantle, with attendant broad implications for chemical reactions, heat transport, and boundary-layer dynamics²⁹. Together with the latest generation of high-spatial-resolution seismic tomography images^{30,31}, these seismic observations and their spatial systematics raise many questions about primary processes in the CMB boundary layer that define challenging research frontiers for experimental and computational geophysics. The present state of knowledge about the CMB region parallels that of the lithospheric boundary layer in the early 1960s, at the dawn of the plate-tectonics revolution. It is reasonable to expect that a concerted effort to quantify the deep boundary layer processes will achieve large advances in our understanding of the dynamic Earth system.

Seismological discoveries

The elastic velocity structure of the lowermost mantle has been studied by seismologists for decades³², yet an important new discovery of some fundamental aspect of the structure in the D'' region has occurred every few years. Although traditional arrival-time analysis procedures for seismic P and S waves continue to lie at the heart of the increasing resolution of seismic tomography models, detailed analysis of waveforms for a variety of secondary phases has played a larger role in revealing high-resolution details of the boundary-layer structure.

Figure 1 illustrates some of the ray paths associated with seismic phases that provide information about the deep mantle structure. P and S waves that graze the D'' region (Fig. 1a), either reflecting off of the core (PcP, ScS) or diffracting along it (P_{diff}, S_{diff}) have been studied for decades, but recent analysis procedures which account for shallow mantle effects are revealing coherent large-scale patterns that augment those detected by travel-time tomography³³. Commonly, differential anomalies such as ScS–S arrival times are used to help isolate lower-mantle contributions. The discovery in the early 1980s of a deep-mantle shear-wave reflector several hundred kilometres above the CMB³⁴ introduced the analysis of reflections and refractions (S_{bc} and S_{cd} in Fig. 1b, and their P-wave counterparts), which are used to map the lateral extent of localized stratification and its depth and material property contrasts. Core-grazing and core-reflected S waves can also be analysed for any shear-wave splitting associated with anisotropic structure near the base of the mantle, although shallow mantle effects need to be accounted for carefully. For examining the seismic properties within tens of kilometres of the CMB, the optimal seismic waves are short-period P reflections (PcP) and S to P conversions (ScP), as well as longer-period S waves that convert at the CMB, transmitting through or diffracting along the CMB as P waves, before they travel back to the surface as S waves (SKS and SPdKS phases in Fig. 1c). PKP phases and their scattered precursors also provide critical information about small-scale structures in D''.

In all cases, the recent availability of large data sets of digital short-period or broadband data from global instrument deployments and global data centres has been critical to the effort to study the deep mantle structure. Based on detailed waveform analysis of the phases mentioned above, constrained by the spatial coverage intrinsic to earthquake and recording-station distributions, recent seismological efforts have mapped out extensive regions of D'' properties that were not even recognized to exist 15 years ago.

D'' ultra-low-velocity layer

The recently discovered low-velocity layer at the base of the mantle is 5–40 km thick, and has P- and S-wave velocity reductions

exceeding 10% (refs 27–29, 35, 36). These velocity variations are much greater than those suggested by travel-time tomography, and this structure is referred to as an ultra-low-velocity zone (ULVZ). The ULVZ is convincingly revealed by delayed and amplified SPdKS phases (relative to SKS) and by precursors to PcP, and has thus far been mapped with the Fresnel-zone (region of coherent wave interaction) coverage shown in Fig. 2a. Although many earlier studies proposed strong velocity decreases in the lowermost mantle from analysis of reflected and diffracted waves, only the recent work provides compelling evidence for the thin basal structure, as well as documenting its lateral variations. Extensive regions of the ULVZ are found beneath the central Pacific Ocean, Alaska, Iceland and Africa. Other areas, indicated in Fig. 2a, have been examined and do not show evidence for this structure, suggesting either that it is not present or that it is thinner than a few kilometres in depth extent, in which case it may be seismically invisible to present methods. In areas such as the central Pacific, the ULVZ shows significant lateral variations in thickness and/or depth extent³⁷. Fluctuations in the data indicate that there is substantial sub-Fresnel-zone (that is, unresolved small-scale) structure in the ULVZ, with it being possible that there are localized

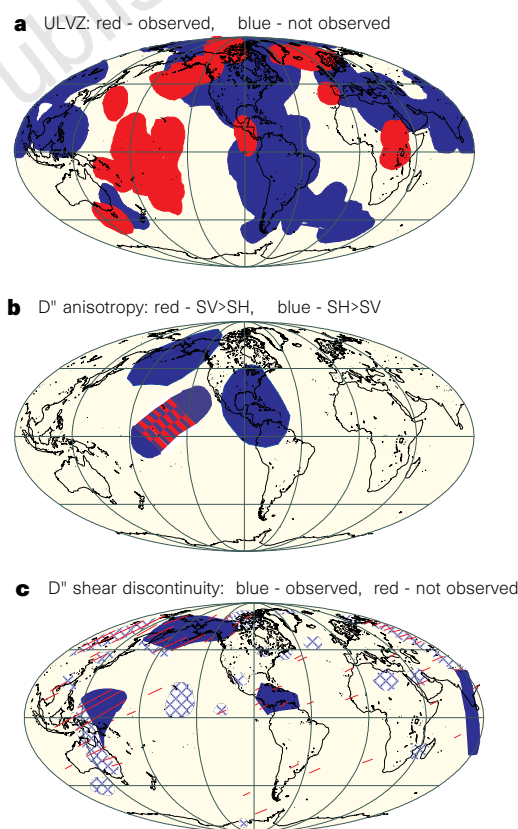


Figure 2 Mollweide projections summarizing spatial patterns of recently determined seismological characteristics of the CMB boundary layer. **a**, Regions with detectable ultra-low-velocity zone (ULVZ) structure are shown in red, and sampled regions that lack evidence of any ULVZ are shown in blue²⁹. **b**, Regions with shear-wave anisotropy in D'' are shown²⁶. Transverse isotropy, caused by horizontal layers with strong velocity contrasts (as indicated schematically) or by hexagonally symmetric minerals with aligned vertical symmetry axes, could account for SH waves traveling faster than SV waves in these regions. The central Pacific region has a mix of no anisotropy and general anisotropy (possibly triclinic, or hexagonal symmetric minerals with variably orientated symmetry axes, or lamellae with variable orientation) with relatively small lateral coherence. **c**, Regions with shear-wave discontinuities near the top of the D'' layer. Several regions show large-scale coherence (solid blue), with mixed areas lacking extra reflections (red stripes), while in some areas the reflection from D'' is highly variable and intermittent (cross-hatched)²¹.

pockets of very acute heterogeneity rather than any type of uniform layer.

Most seismic observations of the ULVZ are directly sensitive to the P-wave velocity decrease, with there being some sensitivity to very large ($>20\%$) S-wave velocity decreases or density increases²⁹. Nonetheless, there is some evidence favouring shear-wave velocity decreases up to three times as strong as the P-wave velocity decrease^{35,36}. The velocity contrasts (the sharpness is not yet well resolved, but is probably less than 5 km) associated with this feature are greater than for any of the main transition-zone discontinuities, and are of the order of those associated with contrasts between subducted slabs and overlying zones of partial melt, or shallow heterogeneities in magma chambers beneath active rifting structures such as mid-ocean ridges. The acute velocity heterogeneity of the ULVZ is a likely source of strong scattering of short-period phases that has long been associated with short-wavelength heterogeneity in D^{''}^{38,39}.

D' anisotropy

Analysis of the polarization of shear waves that graze the D^{''} region has demonstrated the existence of shear-wave anisotropy for certain paths: that is an unexpected discovery given that there is almost no evidence for anisotropic effects on body waves for paths traversing the lower mantle above the D^{''} region^{40,41}. For paths sampling beneath circum-Pacific regions, the observed shear-wave splitting is such that horizontally polarized SH waves arrive ahead of vertically polarized SV waves. Although the data do not fully constrain the geometry of the anisotropy (there is usually very limited azimuthal sampling of any given region), the observations are consistent with transmission through a D^{''} layer with traverse isotropy such as could be produced by hexagonally symmetric crystals with the symmetry axis orientated vertically or by stacks of horizontal thin layers with strong velocity contrasts^{23–25}. The anisotropic zone must extend up at least several hundred kilometres above the CMB.

Under the central Pacific there are rapid spatial variations in shear-wave splitting, sometimes with SH arriving ahead of SV, sometimes with SV arriving ahead of SH, and sometimes with

shear-wave splitting that is not aligned with the great-circle reference frame^{26,42–44} (Fig. 2b). Although azimuthal coverage is again limited, the observations are consistent with patchy regions of general (triclinic) anisotropy or hexagonal symmetry with either vertical or horizontal symmetry axes. Many paths with no evidence for shear-wave splitting are also found. When observed, the anisotropy beneath the Pacific appears to be confined to the lowermost portion of D^{''}, as it tends to increase with distance and, hence, turning-point depth of the seismic rays (Fig. 1a). It is hard to resolve the radial scale of the anisotropic zone owing to strong trade-offs with its lateral extent and ray-path sensitivity to lateral heterogeneity, but the observed rapid fluctuations suggest a thickness of less than 100 km. Although interpretation of deep-mantle shear-wave anisotropy is fraught with uncertainty, as discussed below, it appears that there must be a fundamental difference in structure, deformation style and/or chemistry between the D^{''} region and the overlying mantle.

D'' discontinuity

Over the past 15 years, there has been extensive mapping of the D^{''} P- and S-wave discontinuities involving more than 40 different studies²¹, with the shear-wave structure being extensively detected as shown in Fig. 2c. Many areas have not been analysed owing to the limitations of the source/receiver distribution for analysing reflections from the discontinuity. Where detected, the discontinuity has been successfully modelled as having a strength of 1.5–3% and a depth of 250 ± 100 km above the CMB. The structure appears to be more variable for P waves than for longer-period S waves, and in some places the 'discontinuity' may in fact be a gradational transition zone of up to 50 km thickness rather than an abrupt contrast. In localized patches the discontinuity has either 100–300 km wavelength topography of about 50 km height or corresponding acute lateral gradients in velocity structure, based on the fluctuations in amplitudes and travel times of the reflected phases^{45,46}. The small-scale variability raises the possibility that the discontinuity represents a transition between mantle regions with contrasting physical properties, such as anisotropic fabric or extent of chemical heterogeneity, rather than a conventional rock boundary.

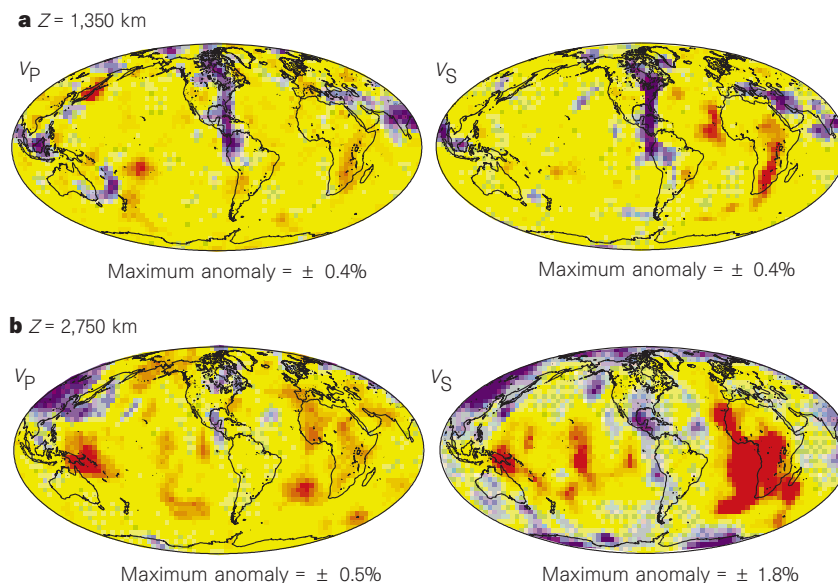


Figure 3 Recent high-resolution global seismic tomography models for P-wave³¹ and S-wave³⁰ velocity heterogeneity near depths (Z) of 1,350 km (top row) and 2,750 km (bottom row). Faster than average velocities are indicated by blues, and slower than average velocities are indicated by reds, with the peak velocity anomaly being shown for each panel (P-wave velocity, V_P , left column; S-wave velocity, V_S , right column). We note the much larger amplitudes of the shear-wave

velocity heterogeneity in the D^{''} layer. The mid-mantle structure appears to be dominated by tabular high-velocity structures, which are generally interpreted to be deep downwellings, possibly associated with slabs penetrating into the lower mantle. These features tend to have coherent over large depth ranges. The D^{''} layer is dominated by large-scale heterogeneities that have limited spatial systematics related to mid-mantle features.

The velocity gradient within the D'' layer beneath the velocity discontinuity appears to be slightly negative, extending over most of the 200–300 km layer thickness²¹. The shear-wave velocity discontinuity is detected by both SH and SV waves, but appears to have weaker strength for SV in the area below Alaska⁴⁷, suggesting that the discontinuity is itself associated with the seismic anisotropy observed in this region^{24,25}. Similarly, there is good evidence for the presence of a shear velocity discontinuity beneath the Caribbean, in the vicinity of the region showing coherent transverse isotropy^{20,23,48}.

Spatial correlations in D'' characteristics

Given the limited spatial coverage of the unusual seismic features in D'' apparent in Fig. 2, establishing any association between these detailed features and the relatively large-scale variations detected throughout the mantle by travel-time tomography can allow global interferences to be made. Heterogeneity patterns of P- and S-wave velocities from recent high-resolution tomographic inversions^{30,31} are shown for the mid-mantle depth of 1,350 km and the mid-D'' depth of 2,750 km in Fig. 3. Of particular interest are possible relations between mid-mantle structures that appear to be dominated by slab-like fast velocity (if these are thermal anomalies, they will be colder, denser, and hence downwelling) heterogeneities (possibly oceanic slabs) and D'' structures. A significant transition in the lateral scale length of heterogeneities between the mid-mantle and the D'' region is apparent in Fig. 3, even allowing for incomplete spatial sampling, and the strength of shear-wave velocity heterogeneity appears to be particularly acute in the D'' region³¹.

The region beneath the central Pacific with the extensive ULVZ in Fig. 2a corresponds to a large-scale low-velocity feature in the tomography models, but there is not a universal correlation between

ULVZ presence and slow D'' structure. The central-Pacific low-velocity region is a relatively stable feature among different tomographic models, and detailed localized studies indicate that the shear-wave velocity has a strongly negative gradient in the lowermost 200–300 km in this general region, with only intermittent evidence for a shear-wave velocity discontinuity⁴⁹. This zone of negative gradient, like the zones of weaker negative gradient beneath regions that have a shear-wave velocity discontinuity, is significantly thicker than the 50–75 km thickness that is often proposed for a thermal boundary layer above the CMB⁵⁰.

The spatial relationship between locations with an observed shear-wave velocity discontinuity and large-scale fast or slow velocity structure in D'' is complex. There is a general tendency for circum-Pacific regions to have higher than average velocities in the data shown in Fig. 3 as well as in lower-resolution tomographic models^{10–12}, but the overall correlation with either P- or S-wave velocity discontinuities is weak²¹. Similarly, any relationship between high-velocity mid-mantle features and the presence of D'' discontinuities is weak: this is perhaps not surprising given that the scale length of lateral heterogeneities appears to shift with depth (Fig. 3).

It is premature to formulate any simplified characterization of how the seismic observations are interrelated on a global basis. Nonetheless, for several of the best-studied regions, there are clear associations that may be summarized as the characteristic dichotomous shear-wave velocity structures in Fig. 4. Regions like those beneath Alaska and the Caribbean have relatively high shear-wave velocities in D'', with a 2–3% SH velocity increase near 200–280 km above the CMB. This velocity increase appears to be weaker for SV, which is consistent with the presence of transverse isotropy that can explain early SH arrivals. This is represented in Fig. 4 by strongly fluctuating horizontal lamellae with high-velocity-contrast materials. Such a laminated structure can account for long-wavelength transverse isotropy⁵¹, but alternative possibilities are discussed below. The velocity gradient throughout the D'' region appears to be negative in these areas, and the strength of lateral heterogeneity reduces with depth. Any ULVZ in these areas is either very intermittent or so thin as to be nondetectable.

In contrast, an area like the central Pacific is characterized by a structure with no significant discontinuity, a strong negative velocity gradient over a depth range of 200 km or more, a thick (up to 40 km) ULVZ, and laterally variable anisotropic structure that appears to be concentrated towards the deeper portion of D''. The central-Pacific anisotropy appears to involve general anisotropy possibly with, in some cases, horizontal symmetry axes such as would be associated with vertical laminations. Although the structures in D'' are undoubtedly not strictly dichotomous, there appear to be large coherent regions with the basic features shown in Fig. 4. These seismological features, and their spatial extent, constitute primary new observational constraints on the thermochemical processes in the lowermost mantle that require interpretations based on our understanding of the physical properties and dynamics of materials at extreme pressures and temperatures. Notably, ideas associated with a simple subsolidus thermal boundary layer are unable to account for most of the features in Fig. 4.

Thermochemical implications

Our knowledge of the complexity of the lowermost mantle has burgeoned over the past 15 years from a zone of decreased compressional- and shear-wave velocity gradients relative to the overlying lower mantle to a zone with pervasive multi-scale heterogeneity, laterally varying seismic discontinuities at depths about 250 km above (and also just above) the CMB, and the presence of variable seismic anisotropy. This challenges computational, theoretical and experimental geophysics to produce viable models for the structure, temperature, flow pattern and chemistry of the lowermost mantle. There has been extensive consideration of possible

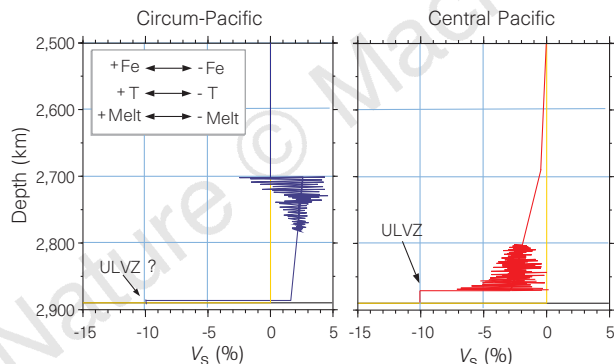


Figure 4 The main classes of D'' shear-wave velocity structures. Schematic shear-wave velocity structures (shown as per cent deviations (δV_s) relative to a standard reference Earth model which has smoothly varying structure in the lower mantle) for circum-Pacific (left) and central Pacific (right) regions. The circum-Pacific regions under Alaska and the Caribbean are well characterized to have relatively abrupt shear velocity increases near 250 km above the CMB, with mild negative gradients throughout the D'' layer. The blue model indicates that these high-velocity regions are relatively cold. These regions also have shear-wave splitting consistent with transverse isotropy, which is represented here by thin horizontal lamellae of strongly varying material properties (possibly involving melt), concentrated near the depth of the discontinuity. Any ULVZ in these regions is so thin as to not be detectable in most cases, and it may not be present. The central Pacific region is characterized by strong negative gradients of shear velocity extending over 200–300 km above the CMB, with little evidence for any discontinuity. The red model indicates that these low-velocity regions are relatively hot. This region has a thick, pronounced ULVZ, with shear velocity decreases that are of the order of 5–30%. There appears to be laterally variable general anisotropy, concentrated towards the base of the D'' region (and possibly in the ULVZ as well). The sense of velocity perturbation with respect to increasing or decreasing Fe concentration, temperature (T) or partial melt is indicated by the inset box.

connections between the CMB and surface observables, with plume-like upwellings rising from a hot CMB thermal boundary layer to generate hotspot volcanism at the surface, and subduction-related cold downwellings producing locally seismically fast (and geochemically anomalous) zones at the CMB. Yet, recent developments in seismology, experimental geophysics and geodynamics indicate that the interaction between hot upwellings, cold downwellings and the CMB are more complex than previously recognized: partial melting, core–mantle chemical interactions, and resultant chemical heterogeneities may each play a role in the dynamic coupling of this layer with the overlying pattern of mantle convection and surface tectonics.

Plumes and the lowermost mantle

The idea that the CMB gives rise to plume-like instabilities and volcanic centres called hotspots is firmly rooted in fluid dynamic experiments and calculations^{5–7,52–55}. Such upwellings require a boundary layer heated from below, and a lowered viscosity within both the plume and a basal layer which serves as a feeder zone for the plume^{52,53,55}. Fluid dynamic stability analysis indicates that the distribution of surface hotspots places a bound of $\geq 10^6$ on the viscosity contrast between such a basal boundary layer and the overlying mantle⁵⁴. As viscosity of minerals is strongly temperature dependent⁵⁶, this lower bound on the boundary layer viscosity contrast could be generated by a temperature jump of 1,000 K, in accord with estimates of the temperature increase across D'' ^{3,4}. Larger changes in viscosity could be generated by the presence of partial melting at the base of the mantle^{29,35,36}; the spatial correlation of the ULVZ with hotspots (Fig. 5 and ref. 57) indicates that deep-seated partial melt may enhance the ascent of a plume to the surface.

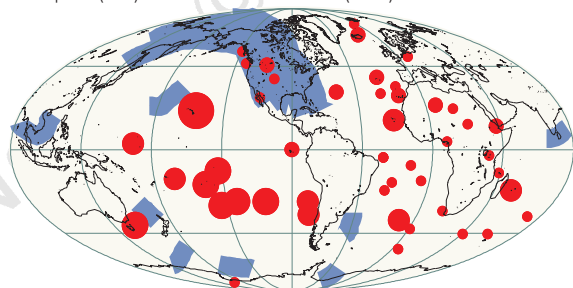
Both petrological modelling of upper-mantle plumes and seismic modelling of lower-mantle variations predict plume excess-temperature anomalies of only about 300 K (refs 58, 59), far less than inferred for plumes ascending from a boundary layer with a

temperature increase near 1,000 K (refs 3, 4). Neither entrainment of surrounding material into the ascending plume⁶⁰, nor diffusional loss of heat from the plume⁶¹, explain this discrepancy. However, imposing a chemical boundary layer at the base of D'' which is denser by at least 5–10% than the overlying mantle decreases the excess temperature of plumes to values in accord with observations⁶². In effect, a chemically stratified boundary layer produces plume extraction from mid-way within the basal thermal boundary layer. Yet the nature of such a chemical boundary layer remains unclear, as do the geodynamic implications of profoundly depressed viscosities in a layer at the base of the mantle. To date, numerical convection simulations have typically not incorporated basal viscosity contrasts which approach a factor of 10^6 or greater: the coupling of large viscosity contrasts with a chemically zoned, laterally heterogeneous boundary layer represents the next generation of geodynamic calculations on CMB properties.

The prevailing view of hotspot generation may thus evolve from one in which plumes originate at the bottom of a thermal boundary layer in a chemically homogeneous mantle, to one in which plumes emerge from a depth some tens of kilometres above the CMB—either within the upper portions of, or above a chemically distinct, partially molten zone (Fig. 6). The zone from which plumes extract material is likely to be both highly inviscid and chemically different from the overlying mantle. The degree and manner in which plumes derived in this way would record a geochemical signature of interaction with the Earth's core are unclear, although hints of such a signature have been reported⁶³.

As plumes have been proposed to have a causal relationship with continental extension and breakup at the surface of the Earth^{53,64}, there is considerable importance in understanding not only the detailed thermochemical dynamics which give rise to plumes, but also the likely evolution of plumes over Earth's history, from the perspective of CMB processes. For example, if the basal portion of D'' has been progressively enriched in relatively dense components over time, a decrease in the excess temperature of such D'' -derived upwellings would have expected over Earth's history: in this sense, the detailed (and obscure⁶⁵) history of hotspot volcanism may provide the only constraints on the properties of the past CMB.

a Hotspots (red) and slabs at the CMB (blue)



b VGP reversal paths

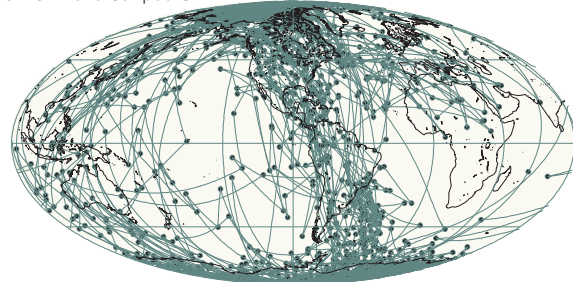


Figure 5 Spatial patterns of large-scale mantle and core dynamical structures. **a**, Map showing spatial distribution of subducted slabs that are expected to have penetrated to the base of the mantle⁶⁴ (blue lines) and buoyancy flux weighted hotspots^{65,64} (red dots). **b**, Map showing virtual geomagnetic pole (VGP) reversal paths⁶³ (green lines). To first order, reversal paths and hotspots appear to be anti-correlated, with reversal paths generally following approximately longitudinal bands in which the ULVZ has not been observed to be present (Fig. 2).

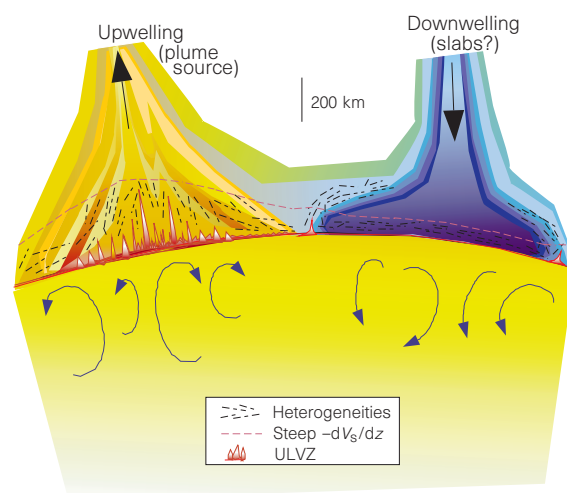
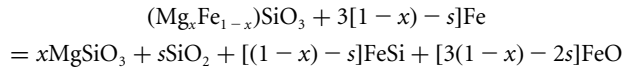


Figure 6 Diagram of possible large-scale CMB structures. The chemical heterogeneities schematically illustrated may arise from partial melting, core–mantle reaction products, slab-associated geochemical heterogeneities, or a combination of these effects. The shape of downwellings at the CMB and the manner in which they interact with CMB structure is unclear (Fig. 3). In the scheme drawn here, the negative buoyancy associated with the downwelling is sufficient to displace much of the pre-existing CMB structure.

Core–mantle chemical reactions

Dynamical simulations indicate that maintaining compositional heterogeneities within D'' requires density increases greater than about 6% to avoid entrainment into the overlying convecting mantle over time^{66,67}. The mechanisms by which such dense material could be introduced into D'' involve either interactions with the underlying core or descent of denser material from the overlying mantle. Chemical reactions between iron alloys and silicates have been shown to occur at high pressures from both experimental and thermodynamic perspectives^{16,68–70}: such reactions are of the form



where $1-x$ is the amount of iron in reacting mantle material, and s is the amount of silica generated by the reaction¹⁶. This reaction of iron with mantle material is uniquely a high-pressure phenomenon, as it does not occur at pressures below about 20–30 GPa (refs 16, 68). The unusual feature of this reaction is that its products—silicates and iron alloys—differ profoundly in their densities and likely thermal and electrical conductivities.

If pervasive core–mantle chemical reactions occur at the CMB, and if FeSi and FeO are able to segregate from their complementary MgSiO₃ and SiO₂, a dense, highly conductive iron-alloy layer could be generated⁷¹. Although the extent of reactions of core material with the mantle remains controversial¹⁷, it is notable that such reactions are markedly enhanced if the reacting silicates are molten⁷⁰. Accordingly, one possible explanation for the manner in which a chemically distinct (and possibly chemically stratified), partially molten layer at the base of the mantle could arise involves varying amounts of reaction between a partially molten layer at the base of the mantle and the outer core, with increasing degrees of reaction as a function of depth. The flow regime within a partially molten basal layer remains obscure²⁶, but any upward flow within such a layer would produce non-diffusion-limited transfer of iron from the outer core into the lowermost mantle^{16,17}, while simultaneously tending to reduce the amount of stratification within the layer.

Chemical heterogeneities from the mantle

Segregation of dense material from the overlying mantle into D'' could occur by downward descent of negatively buoyant liquids or solids. We speculate that partial melt in the lower mantle could be denser than its coexisting solids⁷², particularly if iron partitions preferentially into the liquid, as it does at lower pressures^{73,74}. Indeed, the relatively high ratio of S-wave velocity changes to P-wave velocity changes in the bulk of the lower mantle has been attributed to small degrees of partial melting⁷⁵. If the ULVZ is produced by downward descent of melt from the overlying mantle, then it would be anticipated to be enriched in incompatible (and possibly volatile) elements⁷⁶, and would differ in its composition from mantle that was altered through interaction with the core and melted *in situ* at the CMB. Unfortunately, the properties of melts under deep mantle conditions are largely unconstrained by experiment: both the temperature and composition of the solidus of the lower mantle are unknown at CMB conditions. Therefore, the probable chemistry of the ULVZ remains almost completely conjectural, other than the single observation that material of (Mg, Fe)₂SiO₄ composition appears to undergo eutectic-style melting at CMB pressures and temperatures near 4,300 K (ref. 77). From an experimental viewpoint, the melting behaviour of the lowermost mantle remains a largely unaccessed frontier: both the eutectic chemistry and temperature of likely lower mantle assemblages are ill-constrained, as are the effects of volatile- and iron-enrichment of such material.

Descent of solid material into D'' from the overlying mantle could also alter the chemistry of the lowermost mantle. This may involve

subduction-related processes: chemically differentiated oceanic crustal material might penetrate to D'' providing a distinctive hotspot geochemical reservoir^{8,9}, and possibly generating laminated features that cause seismic anisotropy of the lowermost mantle²³. Although the presence in D'' of material derived from oceanic lithosphere cannot be precluded, the persistence of tabular downwellings to the base of the mantle is suggested in only a few regions^{30,31}. As cold thermal anomalies, slab-related downwellings have long been recognized to be anticorrelated with hotspots⁷⁸ (Fig. 5). Such downwellings, if they impinge on D'', may produce a generally different structure at the base of the mantle than is present in zones of upwellings, and it is this dichotomy which may give rise to the end-member velocity profiles of Fig. 4.

Mechanisms for generating anisotropy in D''

Both downwellings and upwellings should give rise to strong shear flows near the CMB (Fig. 6) that may produce anisotropy in D''. Yet, the very observation of anisotropy in this zone is anomalous relative to the isotropic character of the overlying lower mantle which presumably also has strong shear flows⁴¹. The two primary means for producing D'' anisotropy are through the presence of structural laminations of material with a shear modulus which differs from that of the surrounding mantle, or a shift in the deformation mechanisms producing flow within the boundary layer. Possibilities for producing structural laminations in this region include emplacement of subducted oceanic crust (possibly partially molten), core–mantle reaction products swept up by mantle flow, or orientated inclusions of partial melt^{23,26}. Indeed, orientated melt inclusions are particularly efficient at generating the observed anisotropy of the lowermost mantle²³. If partial melt is ubiquitous and variable in its depth extent in the lowermost mantle^{35,36}, and if strain fields near upwellings and downwellings are sufficiently strong, a lateral smearing of locally elevated partially molten regions could result from the combination of melt and lateral flow at these depths. Such a process is schematically illustrated in Fig. 6, in which the heterogeneous straining of the upper portion of a partially molten zone could generate layered (and anisotropic) structure such as are shown in Fig. 4.

Alternatively, anisotropy might be generated if either the strain history of the lowermost mantle dramatically differs from that of the overlying material, or if there is a shift in ability to acquire anisotropic fabric within the thermal boundary layer at the base of the mantle. Such a shift could be induced by either a change in mineral phase or a change in the deformation regime of the lowermost mantle. There is little evidence for any phase change occurring in any significant lower-mantle mineralogical constituent in the lowermost 300 km of the mantle²⁶. It is possible that the increased temperatures within D'' could drive the deformation mechanism of the deep mantle from a superplastic regime⁷⁹ into a regime of power-law creep. Deformation maps for a range of materials (including olivine) indicate that increasing temperature at moderately high (and constant) shear stress induces such a transition from superplastic flow to power-law creep⁸⁰. The former deformational mechanism is expected to produce notably less preferred orientation than the latter: such a shift in deformational mechanism could produce an apparent onset of anisotropy within D''. Although the virtually unconstrained depth-dependence of shear stress in the lowermost mantle could also play a primary role in modulating the deformational regime and the resultant magnitude of preferred orientation, the existence of anisotropy within D'' could be simply a manifestation of either the hotter temperature in this region or of the flow regime producing stretching of melt inclusions. In either case, the presence of anisotropy is expected to reflect the direction of the local flow and, as such, observations of anisotropy have the prospect of providing a fundamental signature of the direction, depth extent, and lateral variability of the geodynamic flow pattern in the basal boundary layer of Earth's mantle.

Mechanisms for producing D'' discontinuities

The physical and/or chemical changes that produce laterally sporadic discontinuities near 250 km (refs 21, 34) above the CMB remain obscure. The lateral variations in discontinuities (Fig. 2) probably precludes their explanation by a pressure-induced phase transition of a significant lower-mantle constituent, particularly given the lack of observation of such transitions in the main lower-mantle minerals at these pressures^{21,26}. High-pressure transitions in SiO₂ have been proposed to play a role in D'' structure⁸¹, but the manner in which large (and laterally variable) quantities of free silica could be produced in the lowermost mantle is not clear. The possibility that subducted slabs, which are chemically stratified and may maintain a thermally cold signature to depths of the CMB, could produce these discontinuities presents a few difficult, although perhaps not insurmountable, dynamic issues²¹. These include how the thermally cold core (or perhaps basaltic top) of subducted slabs with initial thickness of near 150 km might become essentially horizontally emplaced at depths near 250 km above the CMB. Such a vertically and laterally homogeneous emplacement implies a coherent deposition of the slab above the CMB that is not apparent in the present images of downwellings in this region (Fig. 3). The possible correlation of these discontinuities with observations of D'' anisotropy (Fig. 2) indicates that their shear velocity contrast could be partially associated with the onset of pervasive texturing or preferred alignment of minerals within D''. That the precise cause(s) of the discontinuity 250 km above the CMB remains enigmatic illustrates the broad range of uncertainties which accompany present geodynamic and mineralogic inferences about the basal structure of the mantle: high-pressure experiments on mantle assemblages at CMB pressures remain in their infancy, and material properties required for accurate geodynamic simulations (such as the magnitude of local density and viscosity contrasts) are poorly constrained.

Magnetic-field implications

Compositional variations and partial melting above the CMB have potentially profound implications both for propagation of the geomagnetic field through the mantle and for the manner in which the magnetic field reverses. The magnetic field is produced by fluid flow within the molten metallic outer core, but a heterogeneous high-conductivity layer within the mantle would be expected to produce a screening effect on the core-generated field and a resultant electromagnetic torque on the deep mantle: such a torque may produce observed decadal fluctuations in the Earth's rotation rate⁸². Moreover, the amount of heat flow into the base of the mantle from the core (and thus the flow pattern at the top of the outer core) is controlled by the temperature field within the lowermost mantle: this temperature distribution could be altered by shifts in plume flux. Indeed, a coupling between plume activity, the temperature field in the lowermost mantle, and the frequency of reversals of the geomagnetic field has been proposed to exist^{83,84}.

Qualitatively, there appears to be an anti-correlation between the paths taken by the Earth's pole during magnetic field reversals⁸⁵ and distribution of the ULVZ (Figs 2 and 5). Although the existence of preferred reversal paths is controversial^{86,87}, it is possible that the lowermost mantle influences the manner in which the magnetic field reverses, or may control the surface manifestation of reversals⁸⁴. But the detailed cause of any coupling between the deep mantle and the magnetic field remains uncertain: correlations between reversal paths and deep mantle structure may be generated with the present field strength if the electrical conductivity of the lowermost mantle lies within two orders of magnitude of that of the core⁸⁸. Short-period shifts ("jerks") in the field intensity may indicate, however, that low-conductivity regions exist near the CMB as well⁸⁹. Iron enrichment of the lowermost boundary could certainly enhance the electrical conductivity of the mantle side of the CMB^{16,70,90}. Few data exist on the electrical conductivity of melted or partially molten

lower-mantle assemblages, but those which do exist are consistent with an increased conductivity of molten silicates of between one and two orders of magnitude relative to solid silicates⁹¹. Therefore, an iron-enriched, partially molten zone at the base of the mantle could potentially access electrical conductivities which could influence the manner in which reversals of the geomagnetic field occur, as well as alter the present observed field at the surface of the Earth. Indeed, if the ULVZ is associated with high electrical conductivities, than its lateral variability (Fig. 2) may produce a partially shuttered window through which we view the Earth's magnetic field.

The CMB and deep mantle dynamics

The recognition of laterally heterogeneous layering and likely signatures of vertical and lateral flow near the CMB has dramatically enhanced the complexity of processes attributed to the region. The presence of the largest seismic discontinuity within Earth's mantle at a scant 5–40 km above the CMB, and its association with the onset of partial melting at this depth, has profound implications not only for deep mantle structure, but also potentially for the genesis of hotspots (and thus possibly continental extension at the Earth's surface, as well) and for the properties of the geomagnetic field. The demonstration that the D'' region is the only portion of the lowermost 2,100 km of the Earth's mantle that shows evidence for anisotropy, implying texturing by flow of material at these depths, indicates that it may actually be possible to characterize flow patterns in this dynamic boundary layer.

In spite of these advances, important questions remain. Despite its discovery nearly 15 years ago, the physical origin of the seismic discontinuity lying near 250 km above the CMB remains enigmatic. The manner in which cold downwellings interact with the basal layer of the mantle is unclear, as is the complementary issue of whether slab-associated material ultimately becomes incorporated within the D'' region. In both the amount (and nature) of chemical input derived from downwellings from above, and in the degree to which interaction of D'' with the underlying core occurs, the chemistry of the bottom of the mantle (and particularly of any partially molten zones) continues to be uncertain. As the probable source region for plumes, the temporal evolution of structure at the CMB may be of crucial importance for understanding the geochemical and thermal evolution of our planet. Indeed, it is not known whether the ULVZ has been gradually produced over time by a combination of core interactions and iron- and/or volatile-enriched melts descending from above, or whether it represents the last residue of a terrestrial magma ocean⁹², pinned by negative buoyancy at the base of a thermal boundary layer.

The rapidity with which new structures within D'' have been seismically characterized has (for the moment) outpaced the corresponding interpretive geodynamic and experimental work. Indeed, few studies have been conducted to date which were designed to constrain the nature of the newly discovered seismic structures of this region. Because of the importance of this region, and because of the temporary offset between the observations and interpretive studies—coupled with the accelerating number and enhanced quality of seismic observations—it seems likely that the CMB is about to replace the transition zone between Earth's upper and lower mantle as the region most likely to hold the key to a large number of geophysical problems⁹³. □

T. Lay and Q. Williams are in the Earth Sciences Department, University of California, Santa Cruz, California 95064, USA; E. J. Garnero is in the Seismological Laboratory, University of California, Berkeley, California 94720, USA.

1. Lay, T. Structure of the core-mantle transition zone: A chemical and thermal boundary layer. *Eos* **70**, 54–55, 58–59 (1989).
2. Loper, D. E. & Lay, T. The core-mantle boundary region. *J. Geophys. Res.* **100**, 6397–6420 (1995).
3. Williams, Q. & Jeanloz, R. Melting relations in the iron-sulfur system at high pressures: Implications for the thermal state of the Earth. *J. Geophys. Res.* **95**, 19299–19310 (1990).
4. Boehler, R. Temperatures in the Earth's core from melting-point measurements of iron at high static pressures. *Nature* **363**, 534–536 (1993).

5. Davies, G. F. Ocean bathymetry and mantle convection, 1, Large-scale flow and hotspots. *J. Geophys. Res.* **93**, 10447–10480 (1988).
6. Sleep, N. H. Hotspots and mantle plumes: Some phenomenology. *J. Geophys. Res.* **95**, 6715–6736 (1990).
7. Loper, D. E. Mantle plumes. *Tectonophysics* **187**, 373–384 (1991).
8. Hoffman, A. W. & White, W. M. Mantle plumes from ancient oceanic crust. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
9. Woodhead, J. D., Greenwood, P., Harmon, R. S. & Stoffers, P. Oxygen isotope evidence for recycled crust in the source of EM-type ocean island basalts. *Nature* **362**, 809–813 (1993).
10. Su, W.-J., Woodward, R. & Dziewonski, A. M. Degree 12 model of shear velocity heterogeneity in the mantle. *J. Geophys. Res.* **99**, 6945–6980 (1994).
11. Li, X. D. & Romanowicz, B. Global mantle shear velocity model developed using nonlinear asymptotic coupling theory. *J. Geophys. Res.* **101**, 22245–22272 (1996).
12. Masters, T. G., Johnson, S., Laske, G. & Bolton, H. A shear-velocity model of the mantle. *Phil. Trans. R. Soc. Lond. A* **354**, 1385–1411 (1996).
13. Gubbins, D. Core-mantle interactions. *Tectonophysics* **187**, 385–391 (1991).
14. Bloxham, J. & Jackson, A. Fluid flow near the surface of Earth's outer core. *Rev. Geophys.* **29**, 97–120 (1991).
15. Hide, R., Speith, M. A., Clayton, R. W., Hager, B. H. & Voorhies, C. V. in *Relating Geophysical Structures and Processes: The Jeffreys Volume* (eds Aki, K. & Dmowska, R.) 107–120 (Geophys. Monogr. Ser. 76, Am. Geophys. Union, Washington DC, 1993).
16. Knittle, E. & Jeanloz, R. Earth's core-mantle boundary: Results of experiments at high pressures and temperatures. *Science* **251**, 1438–1443 (1991).
17. Poirier, J.-P. Core-infiltrated mantle and the nature of the D'' layer. *J. Geomagn. Geoelectr.* **45**, 1221–1227 (1993).
18. Lay, T. Seismology of the lower mantle and core-mantle boundary. *Rev. Geophys. Suppl.* 325–328 (1995).
19. Weber, M. et al. in *Seismic Modeling of the Earth's Structure* (eds Boschi, E., Ekström, G. & Morelli, A.) 399–442 (Istit. Naz. di Geophys., Rome, 1996).
20. Ding, X. & Helmberger, D. V. Modeling D'' structure beneath Central America with broadband seismic data. *Phys. Earth Planet. Inter.* **101**, 245–270 (1997).
21. Wyssession, M. E. et al. The D'' discontinuity and its implications. in *The Core-Mantle Boundary* (eds Gurnis, M., Buffett, B. A., Knittle, E. & Wyssession, M.) (Am. Geophys. Union, in the press).
22. Vinnik, L., Romanowicz, B., Le Stunff, Y. & Makeyeva, L. Seismic anisotropy in the D'' layer. *Geophys. Res. Lett.* **22**, 1657–1660 (1995).
23. Kendall, J.-M. & Silver, P. G. Constraints from seismic anisotropy on the nature of the lowermost mantle. *Nature* **381**, 409–412 (1996).
24. Matzel, E., Sen, M. K. & Grand, S. P. Evidence for anisotropy in the deep mantle beneath Alaska. *Geophys. Res. Lett.* **23**, 2417–2420 (1996).
25. Garnero, E. J. & Lay, T. Lateral variations in lowermost mantle shear wave anisotropy beneath the north Pacific and Alaska. *J. Geophys. Res.* **102**, 8121–8135 (1997).
26. Lay, T., Garnero, E. J., Williams, Q., Kellogg, L. & Wyssession, M. E. Seismic wave anisotropy in the D'' region and its implications. in *The Core-Mantle Boundary* (eds Gurnis, M., Buffett, B. A., Knittle, E. & Wyssession, M.) (Am. Geophys. Union, in the press).
27. Garnero, E. J. & Helmberger, D. V. A very slow basal layer underlying large-scale low-velocity anomalies in the lower mantle beneath the Pacific: evidence from core phases. *Phys. Earth Planet. Inter.* **91**, 161–176 (1995).
28. Mori, J. & Helmberger, D. V. Localized boundary layer below the mid-Pacific velocity anomaly identified from a PcP precursor. *J. Geophys. Res.* **100**, 20359–20365 (1995).
29. Garnero, E. J., Revenaugh, J., Williams, Q., Lay, T. & Kellogg, L. Ultra-low velocity zone at the core-mantle boundary. in *The Core-Mantle Boundary* (eds Gurnis, M., Buffett, B. A., Knittle, E. & Wyssession, M.) (Am. Geophys. Union, in the press).
30. Grand, S. P., van der Hilst, R. D. & Widiyantoro, S. Global seismic tomography: A snapshot of convection in the Earth. *GSA Today* **7**, 1–7 (1997).
31. van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. Evidence for deep mantle circulation from global tomography. *Nature* **386**, 578–584 (1997).
32. Young, C. J. & Lay, T. The core mantle boundary. *Annu. Rev. Earth Planet. Sci.* **15**, 25–46 (1987).
33. Wyssession, M. E. Large scale structure at the core-mantle boundary from core-diffracted waves. *Nature* **382**, 244–248 (1996).
34. Lay, T. & Helmberger, D. V. A lower mantle S-wave triplication and the shear velocity structure of D''. *Geophys. J. R. Astron. Soc.* **75**, 799–838 (1983).
35. Williams, Q. & Garnero, E. J. Seismic evidence for partial melt at the base of Earth's mantle. *Science* **273**, 1528–1530 (1996).
36. Revenaugh, J. S. & Meyer, R. Seismic evidence of partial melt within a possibly ubiquitous low velocity layer at the base of the mantle. *Science* **277**, 670–673 (1997).
37. Garnero, E. J. & Helmberger, D. V. Seismic detection of a thin laterally varying boundary layer at the base of the mantle beneath the central-Pacific. *Geophys. Res. Lett.* **23**, 977–980 (1996).
38. Bataille, K., Wu, R. S. & Flatté, S. M. Inhomogeneities near the core-mantle boundary evidenced from scattered waves: A review. *Pure Appl. Geophys.* **132**, 151–174 (1990).
39. Vidal, J. E. & Hedlin, M. A. H. Evidence for partial melt at the core-mantle boundary north of Tonga from the strong scattering of seismic waves. *Nature* **391**, 682–685 (1998).
40. Kaneshima, S. & Silver, P. G. A search for source-side anisotropy. *Geophys. Res. Lett.* **19**, 1049–1052 (1992).
41. Meade, C., Silver, P. G. & Kaneshima, S. Laboratory and seismological observations of lower mantle isotropy. *Geophys. Res. Lett.* **22**, 1293–1296 (1995).
42. Pulliam, J. & Sen, M. K. Seismic anisotropy in the core-mantle transition zone. *Geophys. J. Int.* (in the press).
43. Kendall, J.-M. & Silver, P. G. Investigating causes of D'' anisotropy. in *The Core-Mantle Boundary* (eds Gurnis, M., Buffett, B. A., Knittle, E. & Wyssession, M.) (Am. Geophys. Union, in the press).
44. Ritsema, J., Lay, T., Garnero, E. J. & Benz, H. Seismic anisotropy in the lowermost mantle beneath the Pacific. *Geophys. Res. Lett.* (in the press).
45. Weber, M. P. and S wave reflections from anomalies in the lowermost mantle. *Geophys. J. Int.* **115**, 183–210 (1993).
46. Lay, T., Garnero, E. J., Young, C. J. & Gaherty, J. B. Scale-lengths of shear velocity heterogeneity at the base of the mantle from S wave differential travel times. *J. Geophys. Res.* **102**, 9887–9910 (1997).
47. Lay, T. & Young, C. J. Analysis of seismic SV waves in the core's penumbra. *Geophys. Res. Lett.* **18**, 1373–1376 (1991).
48. Kendall, J.-M. & Nangini, C. Lateral variations in D'' below the Caribbean. *Geophys. Res. Lett.* **23**, 399–402 (1996).
49. Ritsema, J., Garnero, E. & Lay, T. A strongly negative shear velocity gradient and lateral variability in the lowermost mantle beneath the Pacific. *J. Geophys. Res.* **102**, 20395–20411 (1997).
50. Stacey, F. D. & Loper, D. E. The thermal boundary layer interpretation of D'' and its role as a plume source. *Phys. Earth Planet. Inter.* **33**, 45–55 (1983).
51. Backus, G. E. Long-wave elastic anisotropy produced by horizontal layering. *J. Geophys. Res.* **67**, 4427–4440 (1962).
52. Whitehead, J. A. & Luther, D. S. Dynamics of laboratory diapir and plume models. *J. Geophys. Res.* **80**, 705–717 (1975).
53. Duncan, R. A. & Richards, M. A. Hotspots, mantle plumes, flood basalts, and true polar wander. *Rev. Geophys.* **29**, 31–50 (1991).
54. Ribe, N. M. & de Valpine, D. P. The global hotspot distribution and instability of D''. *Geophys. Res. Lett.* **21**, 1507–1510 (1994).
55. Bercovici, D. & Kelly, A. The non-linear initiation of diapirs and plume heads. *Phys. Earth Planet. Inter.* **101**, 119–130 (1997).
56. Karato, S. & Li, P. Diffusion creep in perovskite: Implications for the rheology of the lower mantle. *Science* **255**, 1238–1240 (1992).
57. Williams, Q., Revenaugh, J. & Ganero, E. J. A correlation between ultralow basal velocities in the mantle and hotspots. *Science* (submitted).
58. Watson, S. & McKenzie, D. P. Melt generation by plumes: A study of Hawaiian volcanism. *J. Petrol.* **12**, 501–537 (1991).
59. Nataf, H.-C. & vanDecar, J. Seismological detection of a mantle plume? *Nature* **264**, 115–120 (1993).
60. Farnetani, C. G. & Richards, M. A. Thermal entrainment and melting in mantle plumes. *Earth Planet. Sci. Lett.* **136**, 251–267 (1995).
61. Leitch, A. M., Steinbach, V. & Yuen, D. A. Centerline temperature of mantle plumes in various geometries: Incompressible flow. *J. Geophys. Res.* **101**, 21928–21846 (1996).
62. Farnetani, C. G. Excess temperature of mantle plumes: The role of chemical stratification across D''. *Geophys. Res. Lett.* **24**, 1583–1586 (1997).
63. Walker, R. J., Morgan, J. W. & Horan, M. F. Osmium-187 enrichment in some plumes—evidence for core-mantle interaction. *Science* **269**, 819–822 (1995).
64. Storey, B. C. The role of mantle plumes in continental breakup—case histories from Gondwanaland. *Nature* **377**, 301–308 (1995).
65. Sleep, N. H., Richards, M. A. & Hager, B. H. On the onset of mantle plumes in the presence of preexisting convection. *J. Geophys. Res.* **93**, 7672–7689 (1988).
66. Davies, G. F. & Gurnis, M. Interaction of mantle dregs with convection: lateral heterogeneity at the core-mantle boundary. *Geophys. Res. Lett.* **13**, 1517–1520 (1986).
67. Sleep, N. H. Gradual entrainment of a chemical layer at the base of the mantle by overlying convection. *Geophys. J.* **95**, 437–447 (1988).
68. Goarant, F., Guyot, F., Peyronneau, J. & Poirier, J. P. High pressure and high-temperature reactions between silicates and liquid iron alloys, in the diamond anvil cell, studied by analytical electron microscopy. *J. Geophys. Res.* **97**, 4477–4487 (1992).
69. Song, Y. & Ahrens, T. J. Pressure-temperature range of reactions between liquid iron in the outer core and mantle silicates. *Geophys. Res. Lett.* **21**, 153–156 (1994).
70. Ito, E., Morooka, K., Ujike, O. & Katsura, T. Reactions between molten iron and silicate melts at high pressure: Implications for the chemical evolution of Earth's core. *J. Geophys. Res.* **100**, 5901–5910 (1995).
71. Manga, M. & Jeanloz, R. Implications of a metal-bearing chemical boundary layer in D'' for mantle dynamics. *Geophys. Res. Lett.* **23**, 3091–3094 (1996).
72. Rigden, S. M., Ahrens, T. J. & Stolper, E. M. High-pressure equation of state of molten anorthite and diopside. *J. Geophys. Res.* **94**, 9508–9522 (1989).
73. McFarlane, E., Drake, M. J. & Rubie, D. C. Element partitioning between Mg-perovskite, magnesio-wüstite and silicate melt at conditions of the Earth's mantle. *Geochim. Cosmochim. Acta* **58**, 5161–5172 (1994).
74. Ohtani, E., Kato, T. & Ito, E. Transition metal partitioning between lower mantle and core materials. *Geophys. Res. Lett.* **18**, 85–88 (1991).
75. Duffy, T. & Ahrens, T. J. in *High-Pressure Research: Application to Earth and Planetary Sciences* (eds Syono, Y. & Manghni, M. H.) 197–205 (Terra Scientific, Tokyo, 1992).
76. Knittle, E. The solid/liquid partitioning of major and radiogenic elements at lower mantle pressures: Implications for the core-mantle boundary region. in *The Core-Mantle Boundary* (eds Gurnis, M., Buffett, B. A., Knittle, E. & Wyssession, M.) (Am. Geophys. Union, in the press).
77. Holland, K. G. & Ahrens, T. J. Melting of (Mg, Fe)₂SiO₄ at the core-mantle boundary of the Earth. *Science* **275**, 1623–1625 (1997).
78. Crough, S. T. & Jurdy, D. M. Subducted lithosphere, hot spots and the geoid. *Earth Planet. Sci. Lett.* **48**, 15–22 (1980).
79. Karato, S. I., Zhang, S. & Wenk, H. R. Superplasticity in Earth's lower mantle: Evidence from seismic anisotropy and rock physics. *Science* **270**, 458–461 (1995).
80. Frost, H. J. & Ashby, M. F. *Deformation Mechanism Maps* (Pergamon, Oxford, 1982).
81. Cohen, R. E. in *High-Pressure Research: Application to Earth and Planetary Sciences* (eds Syono, Y. & Manghni, M. H.) 425–431 (Terra Scientific, Tokyo, 1992).
82. Stewart, D. N., Buse, F. H., Walker, K. A. & Gubbins, D. Geomagnetism, Earth rotation and the electrical conductivity of the lower mantle. *Phys. Earth Planet. Inter.* **92**, 199–214 (1995).
83. Larson, R. L. & Olson, P. Mantle plumes control magnetic reversal frequency. *Earth Planet. Sci. Lett.* **107**, 437–447 (1991).
84. Marzocchi, W. & Mulargia, F. The periodicity of geomagnetic reversals. *Phys. Earth Planet. Inter.* **73**, 222–228 (1992).
85. Laj, C., Mazaud, A., Weeks, R., Fuller, M. & Herrero-Berverra, E. Geomagnetic reversal paths. *Nature* **351**, 447–450 (1991).
86. Valet, J. P., Tucholka, P., Courtillot, V. & Meynadier, L. Palaeomagnetic constraints on the geometry of the geomagnetic field during reversals. *Nature* **356**, 400–407 (1992).
87. Hoffman, K. Dipolar reversal states of the geomagnetic field and core-mantle dynamics. *Nature* **359**, 789–794 (1992).
88. Aurnou, J. M., Buttles, J. L., Neumann, G. A. & Olson, P. L. Electromagnetic core-mantle coupling and paleomagnetic reversal paths. *Geophys. Res. Lett.* **23**, 2705–2708 (1996).
89. Courtillot, V. & Valet, J. P. Secular variation of the Earth's magnetic field: From jerks to reversals. *C. R. Acad. Sci.* **320**, 903–922 (1995).
90. Shankland, D. J., Peyronneau, J. & Poirier, J. P. Electrical conductivity of the Earth's lower mantle. *Nature* **366**, 453–455 (1993).
91. Li, X. & Jeanloz, R. Laboratory studies of the electrical conductivity of silicate perovskites at high pressures and temperatures. *J. Geophys. Res.* **95**, 5067–5078 (1990).
92. Tonks, W. B. & Melosh, H. J. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 151–174 (Oxford Univ. Press, New York, 1990).
93. Birch, F. Elasticity and constitution of the Earth's interior. *J. Geophys. Res.* **57**, 227–286 (1952).
94. Lithgow-Bertelloni, C. & Richards, M. A. The dynamics of Cenozoic and Mesozoic plate motions. *Annu. Rev. Earth Planet. Sci.* (in the press).

Acknowledgements. We thank C. Lithgow-Bertelloni, R. van der Hilst and S. Grand for providing access to their models and M. Kendall for providing comments on the manuscript. This work was supported by the US National Science Foundation.

Correspondence should be addressed to T.L. (e-mail: thorne@earthsci.ucsc.edu).

Birth and early evolution of a planetary nebula

Matthew Bobrowsky*, Kailash C. Sahu†, M. Parthasarathy‡ & Pedro García-Lario§

* Orbital Sciences Corporation, 7500 Greenway Center Drive, No. 700, Greenbelt, Maryland 20770, USA

† Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA

‡ Indian Institute of Astrophysics, Koramangala, Bangalore 560034, India

§ ISO Science Operations Centre, Villafranca del Castillo, Apartado de Correos 50727, 28080 Madrid, Spain

The final expulsion of gas by a star as it forms a planetary nebula—the ionized shell of gas often observed surrounding a young white dwarf—is one of the most poorly understood stages of stellar evolution^{1,2}. Such nebulae form extremely rapidly (about 100 years for the ionization) and so the formation process is inherently difficult to observe. Particularly puzzling is how a spherical star can produce a highly asymmetric nebula with collimated outflows. Here we report optical observations of the Stingray nebula^{3,4}, which has become an ionized planetary nebula within the past few decades⁵. We find that the collimated outflows are already evident, and we have identified the nebular structure that focuses the outflows. We have also found a companion star, reinforcing previous suspicions that binary companions play an important role in shaping planetary nebulae and changing the direction of successive outflows⁶.

Our narrow-band images (Fig. 1) show the compact nebula—which we refer to as the ‘Stingray nebula’—that surrounds the central star of He3-1357 (also known as CPD-59°6926, SAO244567 and IRAS17119-5926). In most of the images, the emission appears most strongly concentrated in an ellipse with its major axis subtending 1.6'' from northeast to southwest. (This represents a correction from the orientation that we published previously⁷.) If the ellipse is actually a circular ring viewed obliquely, then our line of sight is inclined from the actual axis of symmetry by 56°. Above and below the ring are bubbles that are blown open at the poles. In earlier images from the Wide Field and Planetary Camera 1 (WFPC1)—obtained before the spherical aberration of the Hubble Space Telescope was corrected—gas could just barely be seen escaping from these bubbles⁷. However, in the new WFPC2 images that we present here, the superior spatial resolution allows us to see that the escaping gas forms a complex system of collimated outflows. (The nomenclature used for the various nebular features is shown in Fig. 2.) In addition, the higher signal-to-noise ratio of the new images reveals that emission extends in all directions well outside the main bubble and ring structures. This is particularly evident in the low-ionization images. In three of these images—[O I] (shown in Fig. 1), [N II] and [S II] (not shown, but similar to [O I])—the ring is much fainter and more mottled than in the other images. In most of the images (including the continuum image), the equatorial ring has dimensions 1.67'' × 0.92''. But in the [O I], [N II] and [S II] images, the minor axis is slightly longer, 1.0'', presenting a visibly larger eccentricity of 0.83 rather than 0.79 in the case of the other images. In these same three images, the southeastern inner collimated outflow appears to be two separate, but closely situated, collimated outflows. The overall shape of the nebula evokes the name ‘Stingray nebula’.

The mass outflow that we observe is consistent with the blue-shifted Si IV (1,394–1,403 Å), C IV (1,548–1,551 Å) and Al III (1,855–1,863 Å) doublets that indicated the presence of a strong stellar wind⁵. The low-ionization images show the strongest nebular emission coming from the collimated outflows close to the holes in the bubbles where gas is escaping. This is consistent with shock

excitation caused by the wind encountering the shell and being focused by it. Although converging flows have been observed previously⁸, and hydrodynamic models have been used to simulate the formation of collimated outflows for almost two decades^{9–14}, this is the first observation that clearly shows the collimating mechanism of bipolar outflows. Although theoretical work on bipolar outflows has often involved asymmetrical mass loss due to rapid stellar rotation or a toroidal magnetic field, or the presence of disks (accretion disks or protostellar disks) to collimate the outflows, the latter process is not the mechanism observed in the Stingray nebula. The outflows in the Stingray are focused by the nebular bubbles which function like nozzles, with the gas leaving through the polar holes. Bipolar structures have been seen both in protoplanetary nebulae (PPN) and in planetary nebulae (PN), but this is the only object that has been followed spectroscopically from its pre-ionized PPN stage^{3,4} through ionization and recognition as a PN⁵. So it is especially interesting that the collimating mechanism is

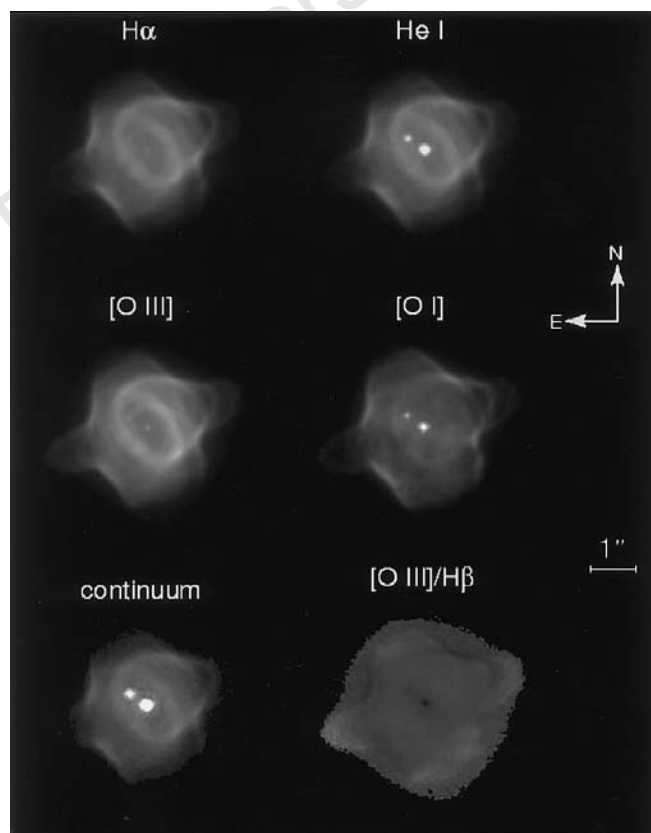


Figure 1 Hubble Space Telescope narrow-band images of He3-1357, the Stingray nebula. The images were obtained in four emission lines (H α , He I 5,876-Å, [O I] 6,300-Å and [O III] 5,007-Å), a continuum range of the spectrum centred at 6,193 Å, and the ratio of [O III] 5,007 Å to H β . Observations were made in March 1996 with WFPC2 using the narrow-band filters F487N, F502N, F588N, F630N and F656N. The continuum exposure was made using filter FQCH4N15 which has a passband of ~44 Å centred at 6,193 Å. Although not shown in this figure, images were also acquired with filters F658N ([N II] 6,583-Å) and F673N ([S II] 6,731-Å) which show the nebula with an appearance almost identical to the [O I] image shown here. The routine scientific data processing included bias, preflash, dark, and flat-field corrections. After removing bad pixels resulting from cosmic rays, both images from the same filter were aligned and averaged. The ratio of the [O III] 5,007-Å to H β emission (bottom right panel) varies across the nebula. Relative to H β , the [O III] emission is lowest ([O III]/H β $\approx$ 2) near the central star, and is also low ([O III]/H β $\approx$ 4) in the denser regions of the bubbles where the gas has been most compressed. The low [O III]/H β ratio in these areas is presumably due to collisional de-excitation occurring where the density is higher. The [O III]/H β ratio is highest (up to 20) in regions of intermediate density, mostly towards the outer parts of the nebula.

clearly visible in an object caught in this brief phase of its evolution. The H β luminosity is similar to the intensity observed⁷ in 1994 from which it was determined that the luminosity of the entire nebular emission is $\sim 5,000$ times the solar luminosity⁵ (assuming a distance of 5.6 kpc to the nebula^{15,16}).

The dynamics in the Stingray nebula are very complex, as evidenced by a number of additional features that are apparent in the HST images. There are two axes of symmetry: one defined by the bright inner ring, and one defined by the polar holes. The images presented here appear to show that the pairs of collimated, bipolar outflows outside the polar holes have different orientations. The position angles of stellar outflows are sometimes seen to vary monotonically with distance from the central star. One example of this occurs in the young planetary nebula He3-1475 (refs 8, 17, 18). This variation is often explained by monotonic precessional motion or a more chaotic "wobbling"^{19,20}, perhaps involving a companion object. Depending on when the various outflows originate relative to the precessional cycle, the position angles of the collimated outflows can show some variation. In the Stingray nebula, the position angles of successive outflow features appear to change with distance from the central star. The inner collimated outflows, which have most recently emerged from the polar holes, are displaced in position angle from that of the polar holes by 3° in an anticlockwise direction. The position angles of the more extended (and therefore older) outer collimated outflows appear displaced by 5° clockwise from the position angles of the inner collimated outflows. If the position of the polar holes controlled the direction of these outflows, then the precessional motion must involve the bubbles containing the polar holes.

The WFPC1 images⁷ did not have sufficient resolution to show a faint companion star embedded in the nebula. But in the higher-resolution images presented here, a companion star is clearly visible. The companion lies $0.4''$ from the central star, at a position angle of 60° , and, at magnitude $V = 17.0 \pm 0.2$ as measured from the continuum filter centred at $6,193 \text{ \AA}$, is 1.6 mag fainter than the central star. (The central star's magnitude of $V = 15.4$ is quite different from previously published magnitudes partly because the luminosity is dropping, as seen from the comparison of ground-based observations^{5,15} between 1985 and 1992, and as expected from the decreasing ultraviolet flux of the central star⁵. Perhaps more importantly, the older, ground-based observations could not resolve the nebula and therefore provided only a combined magnitude of

the star plus the nebula.) The magnitude determination at other wavelengths has higher uncertainty because of the strong nebular emission, but the flux of the companion star is consistent with its being a late-type star at the same distance as the nebula. In considering whether this is truly a companion star, as opposed to a background star, we also made use of the IASG Galaxy Model^{21,22} for the location of the Stingray nebula (galactic longitude $l = 331^\circ$, latitude $b = -12^\circ$). At this location, the probability of finding this star so close ($0.4''$) to the central star is only $\sim 10^{-4}$. The probability is further reduced by a factor of ~ 3 by the additional constraint on its distance as determined from its colour and luminosity. Indeed, other than the central star, no stars this bright are visible anywhere else in the WFPC2 field of view. These considerations leave little doubt that the star is actually a binary companion.

The detection of a binary companion is significant for several reasons. First, the detection of a binary companion is consistent with theoretical expectations. For example, binary nuclei of planetary nebulae should have a double-peaked period distribution, and most should have a period between 30 and 300,000 years⁶. (Below a critical orbital radius, the companion tends to spiral in during a 'common envelope' phase of mass loss. Above the critical orbital radius, mass-loss from the central star causes the orbital radius to increase as the gravitational potential decreases.) The binary nucleus of the Stingray nebula, having a separation of $\sim 2,200 \text{ AU}$, has a period of about 100,000 years—consistent with the theoretical expectations. Second, it has significant implications for the understanding of the early evolution of bipolar structures in planetary nebulae. There are essentially two ways to produce bipolar outflows: one by a density contrast from the pole to the equator²³, and the other by magnetic fields²⁴. In the case of the close binaries, the development of a thick disk in the common envelope phase causes a preferentially bipolar ejection of material. In the case of a wide binary, the companion helps by tidally spinning up the envelope²⁴. The spin-up of the envelope by a companion also reduces the minimum magnetic field required for magnetic effects to be important. When the binary separation is large, as seen here, the effect of the companion star would, at most, produce some small collimating effects²⁵ unless, as described above, its orbit has grown larger owing to mass loss. In the latter case, its effect on the shaping of the nebula may have been more profound. The detection of the binary companion in the case of the Stingray nebula supports the proposed connection between binary central stars and asymmetrical planetary nebulae.

In addition to the overall axisymmetric structure that might be attributable to a companion star, there appears to be a distortion near the northern part of the equatorial ring. (See "spur" in Fig. 2.) This is most easily visible in the H α , H β and [O III] images (Fig. 1), where gas from the ring appears to be drawn towards the companion star. The image shows a number of similar wisps of gas throughout the nebula, so the presence of the spur near the companion star could be coincidental. If, however, it is actually caused by the companion star, that would be notable because, although there has been a fair amount of theoretical work on how a companion star could distort a red-giant envelope, decreasing the pole-to-equator density ratio and thereby producing axial symmetry, the spur may be a case of small-scale distortions of a nebula caused by the binary companion.

The electron temperature was derived using the [O III] line intensity ratios (which are insensitive to the electron density). The resulting temperature is of the order of $1.05 \times 10^4 \text{ K}$, which is typical of planetary nebulae. The electron density can be derived from the H α intensity^{26,27} or the [S II] line ratio⁵, which give a consistent value of $\sim 10^4 \text{ cm}^{-3}$. Given a distance of 5.6 kpc (refs 15, 16), the approximate radius of the nebula is 0.025 pc. A density of 10^4 cm^{-3} would then imply a nebular mass of 0.015 times the solar mass ($0.015 M_\odot$). There is no strong abundance variation in the nebula as seen from the ratio of the line intensities.

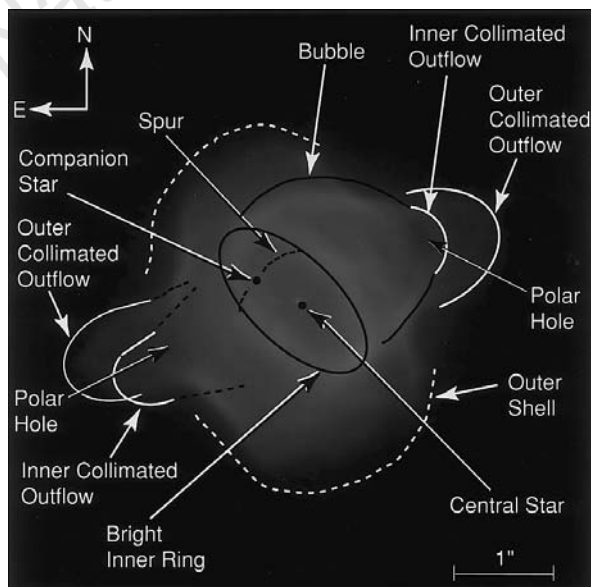


Figure 2 An [O III] 5,007-Å image with various morphological features indicated. See text for details.

A spectrum taken with the International Ultraviolet Explorer (IUE) in 1988 showed⁵ a strong P-Cygni profile (an emission feature with a blue-shifted absorption component), suggesting the presence of a fast wind with terminal velocity of $-3,060 \text{ km s}^{-1}$. The C IV intensity has since decreased monotonically and had fallen below the detection limit by 1994^{28,29}. Our spectrum, taken in 1997, does not show any C IV emission, which confirms that the fast wind has stopped. (The outer shell is the result of the past, slow, red-giant wind.) Because the appearance of multiple collimated outflows implies that the stellar wind is episodic, it is possible that the stellar wind will appear some time again. However, the fact that the ultraviolet flux has also decreased by a factor of 3 within the past eight years²⁸, combined with the decrease of the C IV emission to an undetectable point now, suggests that the central star has, for now, become devoid of nearby circumstellar material. At the same time, the ionization state of the nebula has increased considerably, which suggests that the temperature of the central star is higher.

The simultaneous drop in the ultraviolet flux and the increase in temperature at this stage imply a drop in the total luminosity of the central star. This places the central star of the Stingray nebula at the "knee" of the Hertzsprung–Russell diagram³⁰, just before the star evolves towards lower temperatures and lower luminosities. However, these observations reveal some inadequacy in the current theoretical understanding of how these stars evolve. For the central star to evolve as rapidly as observed here, a stellar mass of $0.8 M_{\odot}$ or more is necessary. However, the luminosity of the central star indicates a core mass of $0.6 M_{\odot}$ or less, for which the evolution is expected to be much slower than observed³¹. The mass loss in the post-AGB (asymptotic giant branch) stage may be the cause of the unexpectedly rapid evolution and may also cause the differences between observations and theory. The observations clearly show that the central star has become a planetary nebula only within the past few decades⁵, and is rapidly evolving to become a DA white dwarf²⁹. Although much work on protoplanetary nebulae and other bipolar objects has been done in an attempt to understand the origin of the axial symmetry and collimated outflows, the Stingray nebula shows how far the nebular structure can develop by the time the nebula becomes ionized. We believe that no other planetary nebula in this phase of its evolution has been previously identified. □

Received 14 August 1997; accepted 29 January 1998.

1. Kwok, S. in *Late Stages of Stellar Evolution* (eds Kwok, S. & Pottasch, S. R.) 321–335 (Kluwer Academic, Dordrecht, 1987).
2. Maddox, J. Making sense of dwarf-star evolution. *Nature* **376**, 15 (1995).
3. Henize, K. G. Observations of southern planetary nebulae. *Astrophys. J. Suppl. Ser.* **14**, 125–153 (1967).
4. Henize, K. G. Observations of southern emission-line stars. *Astrophys. J. Suppl. Ser.* **30**, 391–550 (1976).
5. Parthasarathy, M. et al. SAO 244567: a post-AGB star which has turned into a planetary nebula within the last 40 years. *Astron. Astrophys.* **267**, L19–L22 (1993).
6. Youngelson, L. R., Tutukov, A. V. & Livio, M. The formation of binary and single nuclei of planetary nebulae. *Astrophys. J.* **418**, 794–803 (1993).
7. Bobrowsky, M. Narrowband HST imagery of the young planetary nebula Henize 1357. *Astrophys. J.* **426**, L47–L50 (1994).
8. Borkowski, K. J. et al. Collimation of astrophysical jets: the proto-planetary nebula He3-1475. *Astrophys. J.* **482**, L97–L100 (1997).
9. Cantó, J. A stellar wind model for Herbig-Haro objects. *Astron. Astrophys.* **86**, 327–338 (1980).
10. Morris, M. in *From Miras to Planetary Nebulae* (eds Mennessier, M. O. & Omont, A.) 520–535 (Frontières, Gif-sur-Yvette, 1990).
11. Soker, N. & Livio, M. Disks and jets in planetary nebulae. *Astrophys. J.* **421**, 219–224 (1994).
12. Mellema, G. *Astrophys. Space Sci.* **245**, 239–253 (1996).
13. Frank, A. et al. A mechanism for the production of jets and ansae in planetary nebulae. *Astrophys. J.* **471**, L53–L56 (1996).
14. Frank, A. & Mellema, G. in *Herbig-Haro Flows and the Birth of Low Mass Stars* (eds Reipurth, B. & Bertout, C.) 291–302 (Kluwer Academic, Dordrecht, 1997).
15. Kozok, J. R. Photometric observations of emission B-stars in the southern Milky Way. *Astron. Astrophys. Suppl. Ser.* **61**, 387–405 (1985).
16. Kozok, J. R. Distances, reddening, and distribution of emission B-stars in the galactic centre region $|l| \leq 45^\circ$. *Astron. Astrophys. Suppl. Ser.* **62**, 7–16 (1985).
17. Bobrowsky, M. et al. He3-1475 and its jets. *Astrophys. J.* **446**, L89–L92 (1995).
18. Riera, A. et al. IRAS 17423-1755 [He3-1475]: a massive post-AGB star evolving into the planetary nebula stage? *Astron. Astrophys.* **302**, 137–153 (1995).
19. Livio, M. & Pringle, J. E. The formation of point-symmetric nebulae. *Astrophys. J.* **465**, L55–L56 (1996).
20. Livio, M. & Pringle, J. E. Wobbling accretion disks, jets, and point-symmetric nebulae. *Astrophys. J.* **486**, 835–839 (1997).

21. Ratnatunga, K. U., Bahcall, J. N. & Casertano, S. Kinematic modeling of the galaxy I. The Yale bright star catalog. *Astrophys. J.* **339**, 106–125 (1989).
22. Casertano, S., Ratnatunga, K. U. & Bahcall, J. N. Kinematic modeling of the galaxy. II. Two samples of high proper motion stars. *Astrophys. J.* **357**, 435–452 (1990).
23. Icke, V., Balick, B. & Frank, A. The hydrodynamics of aspherical planetary nebulae II. Numerical modelling of the early evolution. *Astron. Astrophys.* **253**, 224–243 (1992).
24. Livio, M. HST observations of nebular morphologies and their implications. *Space Sci. Rev.* (in the press).
25. Kolesnik, I. G. & Pilyugin, L. S. The influence of a binary nucleus on the structure of planetary nebulae. *Sov. Astron.* **30**, 169–173 (1986).
26. Sahu, K. C. & Desai, J. N. Kinematic structure of NGC 3132: the planetary nebula with a binary nucleus. *Astron. Astrophys.* **161**, 357–362 (1986).
27. Pottasch, S. R. *Planetary Nebulae* 51–70 (Reidel, Dordrecht, 1983).
28. Parthasarathy, M. et al. Fading and variations in the spectrum of the central star of the very young planetary nebula SAO 244567 (Hen 1357). *Astron. Astrophys.* **300**, L25–L28 (1995).
29. Feibelman, W. A. Recent changes in the IUE spectrum of the very young planetary nebula He3-1357 (=SAO 244567). *Astrophys. J.* **443**, 245–248 (1995).
30. Blocker, T. & Schönberner, D. Stellar evolution of low and intermediate-mass stars. III. An application of evolutionary post-AGB models: the variable central star FG Sagittae. *Astron. Astrophys.* **324**, 991–997 (1997).
31. Blocker, T. & Schönberner, D. On the fading of massive AGB remnants. *Astron. Astrophys.* **240**, L11–L14 (1990).

Acknowledgements. We thank T. Heckman for suggestions, M. Livio for discussions, and A. Frank for comments. This work is based on observations with the NASA/ESA Hubble Space Telescope, obtained at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract.

Correspondence and requests for materials should be addressed to M.B. (e-mail: mattb@cta.com).

Unexpected stellar velocity distribution in the warped Galactic disk

R. L. Smart*, R. Drimmel*, M. G. Lattanzi* & J. J. Binney†

* Osservatorio Astronomico di Torino, Pino Torinese, TO 10025, Italy

† Department of Physics, University of Oxford, Oxford OX1 3NP, UK

It is now over 40 years since radio observations of neutral hydrogen revealed¹ the gaseous disk of our Galaxy to be warped. Subsequently, the warp has been detected in the distribution of Galactic dust², molecular clouds³, and luminous stars^{4,5}. Roughly half of all spiral galaxies have similarly warped disks, which suggests that warps are a common and long-lived phenomenon. However, there is still no consensus as to what induces galactic disks to become warped: intergalactic winds, tidal interactions with satellites, magnetic pressure and massive dark haloes have all been proposed as causative agents. Here we use data from the Hipparcos satellite⁶ to determine the small stellar motions in the plane of the sky (proper motions) that should accompany the warp, but which are undetectable in the gas. We find that although the spatial distribution of the stars is in line with previous studies of hydrogen, the velocity distribution has the opposite sign to that expected. Finding a plausible explanation of this result may be the key to solving the long-standing puzzle posed by galactic warps.

A previous attempt to detect warp-induced proper motions⁵ was compromised by the difficulty of accurately detecting a small drift across the sky of a large field of stars. Two special features of the Hipparcos results make it possible to detect accurately such motions in this case: (1) its reference frame was set by distant radio galaxies rather than by a dynamical model of the Solar System; and (2) the measurements over the entire celestial sphere were made by a single instrument. Hipparcos proper motions have random errors $\leq 1 \text{ mas yr}^{-1}$ (ref. 7), accuracies of $\sim 0.1 \text{ mas yr}^{-1}$ (ref. 8) and are inertial to better than 0.25 mas yr^{-1} about any axis⁷.

To study the kinematics of the warp, we received from the Hipparcos Consortium a sample of 2,422 OB stars with positive parallaxes less than 2 mas, and having a Galactic longitude l in the range $(70^\circ\text{--}290^\circ)$. OB stars were chosen because they can be seen to large distances, and are young, so presumably trace the warp in the gaseous disk. By selecting stars with small parallaxes, we focus on distant stars and avoid local, small-scale, structure in the OB stellar

distribution, such as Gould's belt. Unfortunately, the fractional error in the Hipparcos parallactic distances for such distant stars is $\geq 50\%$. Thus Hipparcos indicates only that these are distant stars, and we had to use spectroscopic distances with errors $\sim 20\%$ (ref. 9).

To understand what trends one would expect to find in these data, consider Fig. 1. The Galactic disk is flat to approximately the orbit of the Sun, and then turns up towards the north (Cygnus at $l \approx 90^\circ$), while turning to the south on the opposite side of the Galaxy (Vela at $l \approx 270^\circ$). Hence when we look out from the Sun in the anti-centre direction ($l = 180^\circ$), the band of the Milky Way should slope upwards from large l towards smaller l . More generally, the warp appears as a sinusoidal variation in the distribution of stars on the sky (Fig. 2). In addition, the Galaxy rotates clockwise, so we expect to find stars moving upwards in our sample as they pass through $l = 180^\circ$.

As the real observational situation involves several complexities that are ignored in the simple picture just described, we have compared the data with synthetic data sets (catalogues) derived from a statistical model of the OB-star distribution in phase space. The details of this model are reported elsewhere¹⁰; we mention here that it places the young stars in a warped spiral distribution, models observational errors, and produces catalogues that are defined by the same selection criteria as the actual data set.

The warp is best described in the system of cylindrical coordinates (r, ϕ, z) in which the z -axis runs perpendicular to the plane of the inner Galaxy, r measures galactocentric distance, and the Sun lies at $r_\odot = 8$ kpc on the line $\phi = z = 0$. The azimuth ϕ increases in the direction of Galactic rotation. The average height z of the model disk varies sinusoidally with ϕ :

$$z = h(r)\sin(\phi + \omega_p t) \quad (1)$$

The model warp rotates with angular frequency ω_p , in the opposite direction to that of Galactic rotation for $\omega_p > 0$. We assume that the Sun currently lies precisely on the line of nodes by setting $t = 0$, as the observational signature of the warp is not sensitive to the warp's phase¹¹. As our data extend only a few kiloparsecs from the Sun, covering a small range in ϕ , the stars form a sloped distribution, as seen from the Galactic Centre, in (z, ϕ) at any given r . Consistent with the assumption of a warped disk, the observed slopes (Table 1) are in the correct sense, increase with r , and indicate that the warp begins within the orbit of the Sun, in agreement with recent infrared

observations¹². From these data, and consistent with neutral hydrogen maps¹³, we find the height function $h(r)$ of equation (1) to be

$$h(r) = 0.067(r - 6.5)^2 \text{ kpc} \quad (2)$$

for $r > 6.5$ kpc. We note that h does not depend on t , and ω_p does not depend on r , reflecting our assumption that the warp is long-lived, that is, lasting longer than a few Galactic rotation times.

We now turn to the proper motions. At time t the azimuth of any star is given by $\phi = \phi_0 + \Omega(r)t$, where ϕ_0 is a constant and $\Omega(r)$ is the circular frequency. Differentiating equation (1), we find the predicted vertical velocity is given by:

$$V_z(r) \equiv \frac{dz}{dt} = [\Omega(r) + \omega_p]h(r)\cos(\phi) \quad (3)$$

This equation describes the mean systematic vertical velocity that must be associated with any long-lived warp of the form of equation (1).

From the Hipparcos proper motions we calculate an observed vertical velocity for each star, corrected for Galactic rotation and solar motion (found¹⁰ to be $(U, V, W)_\odot = (9, 5, 7) \text{ km s}^{-1}$). In doing so we have assumed circular rotation and a linear rotation curve, adopting a circular velocity of $v_c = 220 \text{ km s}^{-1}$ at the Sun and a slope $dv_c/dr = -3 \text{ km s}^{-1} \text{ kpc}^{-1}$. The points in Fig. 3 show median values of this observed velocity binned in galactocentric distance, r . If equation (3) were valid, the medians would scatter about the curve defined by the difference between the values it predicts for $V_z(r)$ and $V_z(r_\odot)$. Figure 3a shows two such curves for $\omega_p = 0$ and $-20 \text{ km s}^{-1} \text{ kpc}^{-1}$. It is evident that $\omega_p > -20 \text{ km s}^{-1} \text{ kpc}^{-1}$ does not predict the sign of the observations correctly, and even for large negative values of ω_p the model is a poor

Table 1 Slopes of z versus ϕ for synthetic and observed catalogues

Bin (kpc)	No warp	Warp	Data
$r < 8.5$	-0.716 ± 1.31	2.60 ± 1.40	2.74
$8.0 < r < 9.0$	-0.192 ± 1.34	4.10 ± 1.33	5.06
$8.5 < r < 9.5$	0.611 ± 1.06	6.84 ± 2.65	6.74

Slopes are given in units of pc per degree. The slopes are for stars with $m_V \leq 7.5$, the completeness limit of this Hipparcos sample. For the warped and non-warped cases (columns 2 and 3), the slopes are means and standard deviations of 30 catalogues generated from a statistical model of the OB stellar distribution. The model for the warped case uses equations (1) and (2).

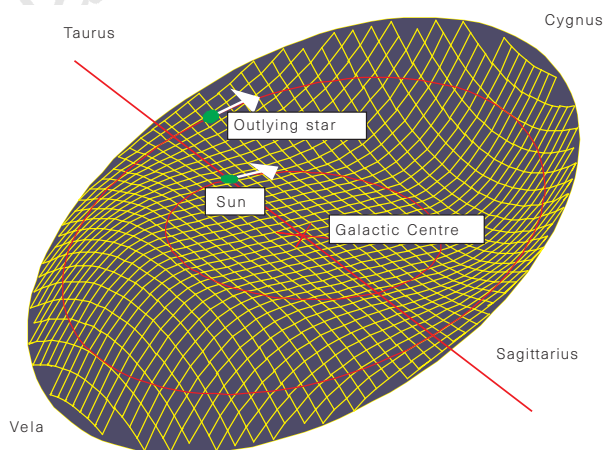


Figure 1 The disk of the Galaxy based on H I observations. The vertical axis is exaggerated by a factor of 10. The arrows show the motion of the Sun and an outlying star on their orbits in the Galaxy. We note that the outlying star has an upward motion as seen from the Sun. Directions in the sky, as seen from the Earth, are indicated by the constellation names. The red line indicates the locus of zero vertical displacement for the warped disk.

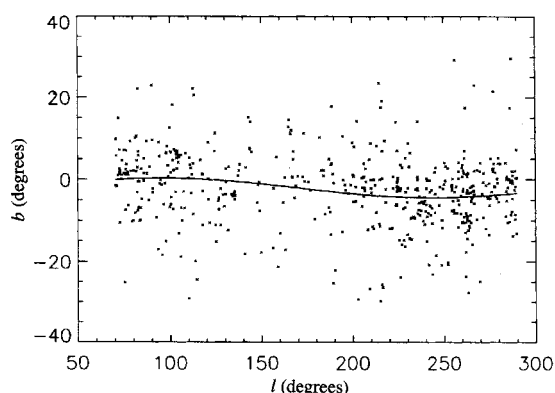


Figure 2 The distribution of OB stars in Galactic coordinates to the completeness limit, $m_V < 7.5$. The curved line is a sine fit to the data. This shape results from stars in the first and fourth quadrant being on average closer to the Galactic Centre, and therefore lower in the warp structure.

fit to the data. This result is surprising because all current theories of warping predict that $\omega_p > 0$ (ref. 14). Figure 3b shows the individual velocities; the scatter within 9.5 kpc is dominated by the stellar peculiar velocities, whereas outside 9.5 kpc the errors begin to contribute significantly. Despite the large effects of errors and peculiar velocities it can still be seen that the data distribution has a systematic downward trend as indicated by the medians in Fig. 3a. All of the simulated kinematic warps using equations (1) and (3) show an upward trend contrary to the observations.

Are the observations flawed? There are three ways in which the data could be misleading: (1) there are systematic errors of order 2 mas yr^{-1} in the measured proper motions; (2) the sample is somehow biased towards negative values of latitude proper motions; and (3) we are seeing something other than the warp. We believe that (1) is excluded by the demonstration that internal systematic errors are smaller than 0.1 mas yr^{-1} . A comparison of the complete sample (V -band magnitude $m_V < 8$) with the selected sample ($8 < m_V < 13$) shows no evidence for biasing. There is a chance that many of the stars form a moving group; however, from

an analysis of the data at $8.5 < r < 9.5$ (where r is in units of kpc), such a group would have to be larger than 2 kpc in extent. Possibility (3) is suggested spatially by neutral hydrogen observations¹⁵. Deviations of $\sim 50 \text{ pc}$ from the plane are seen within the orbit of the Sun. These deviations are certainly affecting our determination of the warp parameters, but the change in z in the last distance bin is $\sim 200 \text{ pc}$, twice the maximum variation within the Sun's orbit.

Although it depends on little more than the assumption that the warp's shape lasts longer than a Galactic rotation period, could equation (3) be invalid? Equations analogous to this are found to be compatible with observations of hydrogen in other warped galaxies, which the Milky Way presumably resembles. We also cannot exclude the possibility that, although the disk has recently been bent into the shape of a warp, it will soon look very different. A possible disturbing agent is the Sagittarius dwarf galaxy¹⁶, but numerical simulations of this interaction are required before this suggestion can be considered plausible. If, rather, the Galactic warp is long-lived, then large systematic motions must be present within the Galactic disk to account for the observations. However, given that no corresponding spatial structures have been identified, such systematic motions at present lack an explanation. We are therefore obliged to conclude that there is currently no convincing interpretation of the implications of Hipparcos data for the dynamics of the warp in the Galactic disk. $\square$

Received 15 August 1997; accepted 5 January 1998.

1. Kerr, F. J. A Magellanic effect on the galaxy. *Astron. J.* **62**, 93 (1957).
2. Freudenreich, H. T. et al. DIRBE evidence for a warp in the interstellar dust layer and stellar disk of the Galaxy. *Astron. J.* **429**, L69–L72 (1994).
3. Wouterloot, J. G. A., Brand, J., Burton, W. B. & Kwee, K. K. IRAS sources beyond the solar circle. II. Distribution in the Galactic warp. *Astron. Astrophys.* **230**, 21–36 (1990).
4. Djorgovski, S. & Sosin, C. The warp of the Galactic stellar disk detected in IRAS source counts. *Astrophys. J.* **341**, L13–L16 (1989).
5. Miyamoto, M., Yoshizawa, M. & Suzuki, S. An optical warp of the galaxy. *Astron. Astrophys.* **194**, 107–115 (1988).
6. Perryman, M. A. C. HIPPARCOS—astrometry from space. *Nature* **340**, 111–113 (1989).
7. Perryman, M. A. C. *The Hipparcos and Tycho Catalogues* (Tech. Rep. SP-1200, ESA, 1997).
8. Bucciarelli, B. in *Proc. HIPPARCOS Venices '97* (ed. Battick, B.) 45–48 (SP-402, ESA Noordwijk, 1997).
9. Smart, R. L., Drimmel, R. & Lattanzi, M. G. in *Proc. HIPPARCOS Venice '97 Proc.* (ed. Battick, B.) 715–720 (SP-402, ESA Noordwijk, 1997).
10. Drimmel, R., Smart, R. L. & Lattanzi, M. G. in *Proc. HIPPARCOS Venice '97* (ed. Battick, B.) (SP-402, ESA Noordwijk, 1997).
11. Smart, R. L. & Lattanzi, M. G. The Galactic warp in young star proper motions. *Astron. Astrophys.* **314**, 104–107 (1996).
12. Freudenreich, H. T. The shape and color of the galactic disk. *Astron. J.* **468**, 663–678 (1996).
13. Dipsas, A. & Savage, B. D. Neutral hydrogen in the outer galaxy. *Astrophys. J.* **377**, 126–140 (1990).
14. Binney, J. J. Warps. *Annu. Rev. Astron. Astrophys.* **30**, 51–74 (1992).
15. Malhotra, S. The vertical distribution and kinematics of HI and mass models of the galactic disk. *Astrophys. J.* **448**, 138–148 (1995).
16. Ibata, R. A., Gilmore, G. & Irwin, M. J. A dwarf satellite galaxy in sagittarius. *Nature* **370**, 194–196 (1994).

Correspondence and requests for materials should be addressed to R.L.S. (e-mail: smart@to.astro.it).

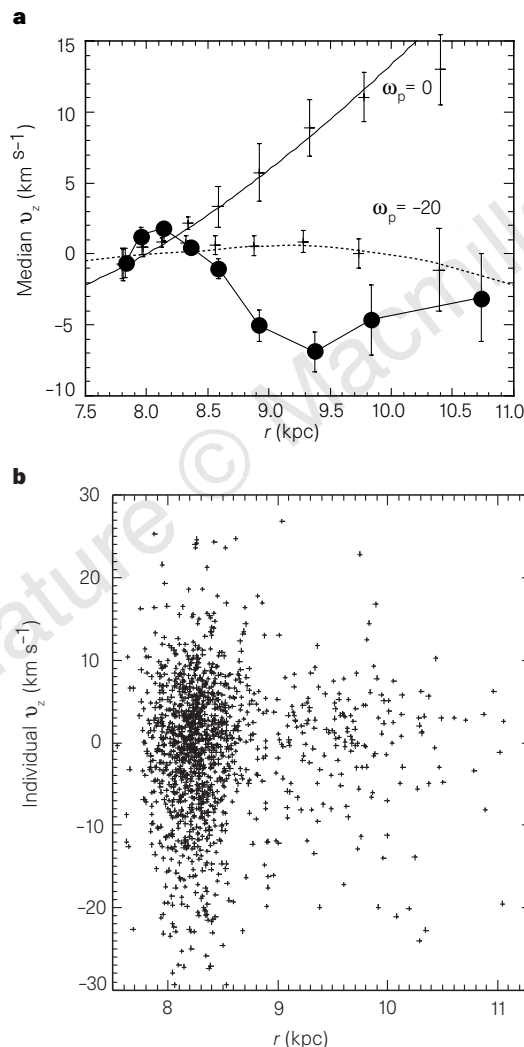


Figure 3 The systematic vertical velocities as a function of Galactocentric distance. **a**, The filled circles show medians of the observed vertical velocity, the error bars indicating their uncertainty. The curves show predicted $V_z(r) - V_z(r_\odot)$ using equation (3), for $\omega_p = 0$ (full line) and $\omega_p = -20 \text{ km s}^{-1} \text{ kpc}^{-1}$ (dashed line). Their error bars represent one standard deviation of medians from 30 simulated catalogues for each set of warp parameters. **b**, Plot of the derived individual velocities which shows the large scatter due to the effects of peculiar velocities, especially within 9.5 kpc where they dominate.

Pairing of charge-ordered stripes in (La,Ca)MnO₃

S. Mori*, C. H. Chen* & S.-W. Cheong*†

* Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

† Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08855, USA

The propensity of systems of charge and spin to form, under certain conditions, 'stripe' phases has recently attracted much attention, as it has been suggested that dynamically fluctuating stripe phases may be of central importance for an understanding of the physics of high-temperature superconductors^{1–5}. A related phenomenon—static charge stripes—characterizes⁶ the insulating antiferromagnetic ground state of the manganese oxides, a class of materials which (like the copper oxide superconductors) have a perovskite structure, and are notable for their extraordinary

electronic and magnetic properties, such as colossal magnetoresistance and charge ordering^{7,8}. Here we report a different pattern of charge localization in the charge-ordered phase of the manganese oxide $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ ($x \geq 0.5$). This pattern takes the form of extremely stable pairs of Mn^{3+}O_6 stripes, with associated large lattice contractions (due to the Jahn–Teller effect), separated periodically by stripes of non-distorted Mn^{4+}O_6 octahedra. These periodicities, which adopt integer values between 2 and 5 times the lattice parameter of the orthorhombic unit cell, correspond to the commensurate carrier concentrations ($x = 1/2, 2/3, 3/4$ and $4/5$): for other values of x , the pattern of charge ordering is a mixture of the two adjacent commensurate configurations. These paired Jahn–Teller stripes appear therefore to be the fundamental building blocks of the charge-ordered state in the manganese oxides, and so may be expected to have profound implications for the magnetic and transport properties of these materials.

Experimental details have been described elsewhere⁶. Thin samples ($<1,000 \text{ \AA}$ thick) of $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ ($x \geq 0.5$) suitable for electron diffraction and electron microscopy studies were prepared from bulk polycrystalline materials with a typical grain size of about a few micrometres. In Fig. 1a we show a $[001]$ zone-axis high-resolution lattice image for $x = 2/3$ taken at a temperature of 95 K. (We note that all the high-resolution lattice images shown here are obtained from $[001]$ -orientated grains with a large objective aperture covering scattering angles slightly larger than the (220) Bragg reflections which correspond to a spacing of $1.9 \text{ \AA}$ in real space. The images are therefore dominated by the lattice distortions rather than the charge transfer which dominates only at small scattering angles.) The electron-diffraction patterns from which the high-resolution images were obtained have been published previously^{6,9,10}. Electron diffraction shows that the charge ordering is commensurate with a

wavevector of $(2\pi/a_0) (1/3, 0, 0)$ and $a_0 = 5.5 \text{ \AA}$, as previously reported^{6,9}. The lattice fringes corresponding to $a_0 = 5.5 \text{ \AA}$ planar spacing are readily resolved in Fig. 1a. The most striking feature of Fig. 1a is the presence of doublet lattice-fringes with $5.5\text{-\AA}$ spacing in each of the dark stripes. The regular array of the paired dark fringes exhibits a $3a_0 (=16.5 \text{ \AA})$ periodicity. From the analysis of the high-resolution lattice images, we conclude that the enhanced contrast of the paired dark fringes is mainly due to the strain induced by the Jahn–Teller distortion of the Mn^{3+}O_6 octahedra. The high-resolution lattice image, therefore, reveals that the charge-ordered structure is constituted of two heavily Jahn–Teller-distorted diagonal Mn^{3+}O_6 stripes (JTS) which always appear in pairs. The much less (or non-) distorted Mn^{4+}O_6 stripes exhibit a much lighter contrast. The high-resolution lattice image shown in Fig. 1a is fully consistent with the schematic model shown in Fig. 1b for the carrier concentration of $x = 2/3$. This model for $x = 2/3$ actually can be derived from the ordering scheme for $x = 1/2$ (see Fig. 3b)¹¹: every third vertical Mn^{3+} stripe in the $x = 1/2$ scheme is replaced by Mn^{4+} rows and the d_{z^2} orbitals of subsequent Mn^{3+} stripes are re-orientated to form a $3a_0$ periodic array. We see immediately that the Mn^{3+} stripes separated by $5.5 \text{ \AA}$ (a -axis lattice parameter) form pairs in this new configuration, and that the new periodicity of this charge-ordered state is now $3a_0 (=16.5 \text{ \AA})$. The strong pairing tendency of the JTS is nicely demonstrated in an inverted intensity trace along the direction perpendicular to the stripes (Fig. 1c), where the pairs of the strongest peaks are due to Mn^{3+}O_6 stripes and the weaker peaks and valleys are identified as Mn^{4+}O_6 stripes. The same spatial scale is used in Fig. 1b and c to clarify the association of the peaks/valleys with the $\text{Mn}^{3+}/\text{Mn}^{4+}$ stripes. The $3a_0$ charge-ordering periodicity between pairs of the Mn^{3+}O_6 stripes is evident, and every Mn^{3+} and Mn^{4+} row is clearly resolved in the high-resolution lattice image.

To verify that pairs of JTS are also fundamental building blocks for other commensurate carrier concentrations, we show in Fig. 2a a similar high-resolution lattice image of the charge-ordered state from a sample with $x = 3/4$. Pairs of dark stripes, with contrast enhanced by the Jahn–Teller distortion of the Mn^{3+}O_6 octahedra, are again evident and the distance between the two adjacent pairs in the periodic array is now $22 \text{ \AA}$ corresponding to a spacing of $4a_0$, as shown by diffraction⁶. An inverted intensity scan (Fig. 2b) again

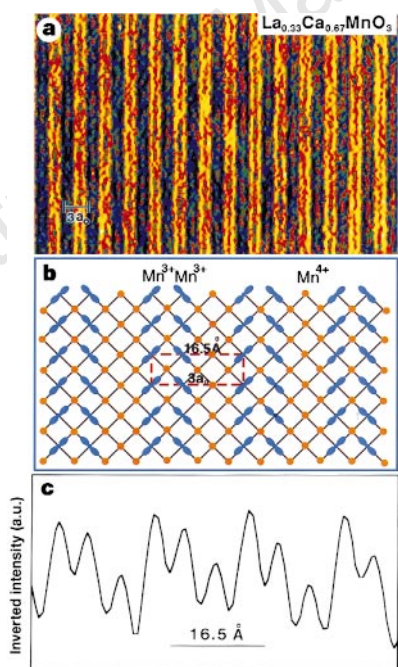


Figure 1 Pairing of charge-ordered stripes in $\text{La}_{0.33}\text{Ca}_{0.67}\text{MnO}_3$. **a**, High-resolution lattice image obtained at 95 K showing $3a_0$ pairing of JTS. **b**, Schematic model in the a - b plane showing the pairing and orbital ordering of Mn^{3+} JTS in blue, and the Mn^{4+} ions in orange. **c**, Inverted intensity scan from a selected area in **a** (a.u., arbitrary units). Pairs of the strongest peaks are identified as Mn^{3+}O_6 JTS and the weaker peaks and valleys as Mn^{4+}O_6 stripes. The same horizontal distance scale is used in **b** and **c** to facilitate the comparison/identification.

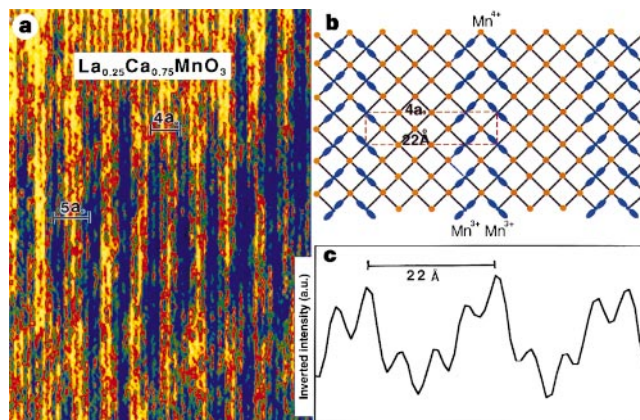


Figure 2 Pairing of charge-ordered stripes in $\text{La}_{0.25}\text{Ca}_{0.75}\text{MnO}_3$. **a**, High-resolution lattice image obtained at 95 K showing pairing of JTS with $4a_0$ periodicity; **b**, schematic model showing three paired JTS; **c**, an inverted intensity scan with the same horizontal scale. A stacking fault of $5a_0$ paired JTS is also indicated in **a**.

demonstrates the strong pairing tendency of the Mn^{3+} stripes in this case. A schematic diagram showing the charge/orbital ordering of JTS is also depicted in Fig. 2c directly below Fig. 2b with the same spatial scale. The configuration of JTS in Fig. 2b can be derived from Fig. 1b by simply increasing the separation between the paired Mn^{3+} stripes from $3a_0$ to $4a_0$. Counting of charges in this case leads to a $\text{Mn}^{3+}/\text{Mn}^{4+}$ ratio of 1:3, as expected from the carrier concentration of $x = 3/4$. In samples with $x = 3/4$, we have occasionally observed the JTS configurations with periodicities of $3a_0$ and $5a_0$ as defect structures. Such a 'defect' with $5a_0$ paired-JTS configuration is also marked in the high-resolution lattice image shown in Fig. 2a.

The coherence length of the $4a_0$ JTS is $\geq 200 \text{ \AA}$ in the transverse directions of the stripes. The $3a_0$ JTS observed for $x = 2/3$ has a transverse coherence length exceeding $500 \text{ \AA}$. The longitudinal coherence length is much greater than the transverse direction and may exceed $1,000 \text{ \AA}$. The coherence length along the c -axis is estimated to be at least of the same order as the sample thickness which is typically $< 1,000 \text{ \AA}$. Hence, the Jahn–Teller stripes we are dealing with here are in fact Jahn–Teller 'sheets' projected in a direction parallel to the incident electron beam. The strong phase correlation along the longitudinal direction of the stripes indicates the importance of the long-range nature of strains in the paired JTS. It is energetically unfavourable to form defects owing to phase shift of the paired stripes. Our studies of the samples with $x = 4/5$ reveal similar JTS configuration as $x = 3/4$, but with a reduced overall contrast and coherence length as suggested from the weaker and broader superlattice peaks observed in the diffraction patterns. The $5a_0$ JTS configuration occurs slightly more often in this case than in $x = 3/4$, but the JTS configuration is still dominated by the presence of $4a_0$ JTS. This may be due to reduced correlation of $5a_0$ JTS.

A crucial test of paired JTS as a generic property of the charge-ordered state in manganites occurs for $x = 1/2$ where previous diffraction studies have demonstrated a $2a_0$ charge modulation^{10,12}. Our high-resolution lattice images obtained from the $x = 1/2$ sample at low temperatures indeed show the pair formation of JTS with a similar contrast characteristic to that shown in Figs 1a and 2a. The pairing of JTS for $x = 1/2$ is best demonstrated in an

inverted intensity scan (Fig. 3a). It is clear that strong peaks, corresponding to Jahn–Teller-distorted Mn^{3+} stripes, form pairs separated by deep valleys (Mn^{4+} stripes). From the schematic charge-ordering model of $x = 1/2$ shown in Fig. 3b, we see that the paired Mn^{3+} stripes have d_{z^2} orbitals orientated perpendicularly to each other, just as in the paired stripes in $x = 2/3$ and $4/5$.

The stability of the paired JTS is best revealed through a quantitative analysis of the intensity scans shown Figs 1c, 2c and 3a. We find that the separation of paired JTS is actually smaller than the normal undistorted distance of a_0 ($= 5.5 \text{ \AA}$) and that this contraction is compensated by the dilation of non-paired stripes. Ordering of JTS in $x = 1/2$ is characterized by two distinct spacings, S_1 for paired stripes and S_2 for non-paired stripes as indicated in Fig. 3a. A histogram of S_1 and S_2 (with a bin-width of $\sim 0.16 \text{ \AA}$) is shown in Fig. 3c; this unambiguously demonstrates the contraction (dilation) of the paired (non-paired) stripes with respect to the undistorted separation of $a_0 = 5.5 \text{ \AA}$. Similar histograms were also obtained for paired JTS in $x = 2/3$ and $3/4$. The actual amount of contraction (dilation) between paired JTS tends to vary locally from stripe to stripe, and the averaged contraction (dilation) is $\sim 0.8 \text{ \AA}$. The displacement of each Jahn–Teller stripe ($\sim 0.4 \text{ \AA}$) from its normal position accounts for $\sim 7\%$ of a_0 , which is comparable to the $\sim 10\%$ Jahn–Teller distortion associated with the Mn^{3+}O_6 octahedra in LaMnO_3 (ref. 13). The local fluctuation of lattice distortions has also been observed in recent neutron and X-ray diffraction experiments^{14,15}.

We now consider what will happen to the charge-ordered state for samples with arbitrary carrier concentrations. Our experimental results for such cases show that charge ordering appears as proper mixtures of the two adjacent distinct commensurate configurations according to the lever rule. For example, a sample with $x = 5/8$, between $1/2$ and $2/3$, in the charge-ordered state is found to exhibit a fine mixture of $25\% 2a_0$ and $75\% 3a_0$ paired JTS configurations. The minority $2a_0$ configuration, however, is not phase-separated into a sizeable area, but instead, appears as incoherent stacking-fault defects in the otherwise perfect $3a_0$ configuration. The frequency of the appearance of $2a_0$ JTS to $3a_0$ JTS is governed completely by the lever rule. The fine mixture of these distinct JTS configurations is the origin of the incommensurate charge-ordered state of $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ for $x > 0.5$.

The physical origin of pairing of the Mn^{3+} stripes is caused by the minimization of the strain energy associated with the Jahn–Teller-distorted Mn^{3+}O_6 octahedra. Although the detailed lattice distortion pattern is not known for the paired JTS configurations, the charge/orbital ordering pattern in the immediate neighbourhood of the paired Mn^{3+} stripes is very similar to $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ (see Fig. 3b) suggesting a similar lattice distortions in these two cases. In the case of $x = 1/2$, most of the lattice distortions are concentrated on the Jahn–Teller-distorted Mn^{3+}O_6 octahedra and almost no distortions occur on the Mn^{4+}O_6 octahedra¹². Each non-paired Mn^{3+} stripe will cause significant lattice distortion in its immediate neighbourhood and the overall strain energy can be lowered by forming periodic array of pairs of Mn^{3+} stripes separated by undistorted regions of Mn^{4+} ions. The gain in lowering the elastic energy of JTS pairs obviously outweighs the cost of Coulomb energy. The size of the strain-free Mn^{4+} regions should be as large as possible (but subject to the constraint imposed by the carrier concentration) to minimize the total strain energy. Experimentally, grouping of more than two Mn^{3+} stripes is not observed. In some respects, the physics for the formation of JTS pairs is similar to the discommensuration theory of incommensurate phases proposed by McMillan¹⁶, namely, the concentration of lattice strains in the narrow discommensurations separated by the strain-free commensurate regions.

Although the phase correlation is strong along the longitudinal direction of the paired stripes, amplitude (contrast) fluctuations along the stripes (on a length scale less than a few hundred

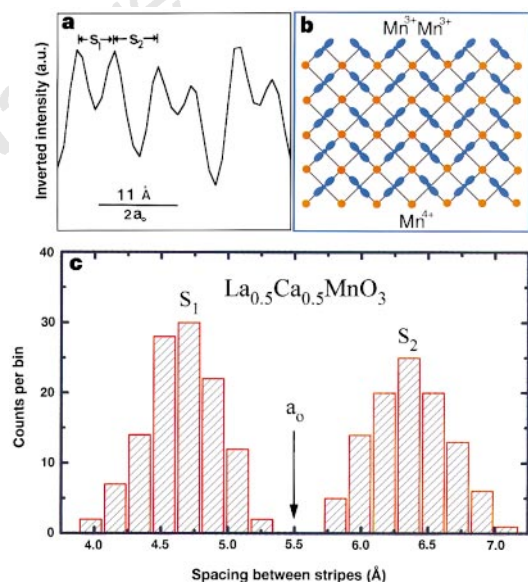


Figure 3 Pairing and lattice distortions of charge-ordered stripes in $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$. Inverted intensity scan (a) and schematic (b) showing ordering of JTS. A histogram of two distinct spacings, S_1 for paired stripes and S_2 for non-paired stripes, is shown in c which clearly demonstrates contractions for S_1 and dilations for S_2 from the non-distorted average spacing of $5.5 \text{ \AA}$ ($= a_0$).

ångströms) are commonly observed. The fluctuations arise from local distortion variations of the $M^{3+}O_6$ octahedra. The fluctuation length scale seems to decrease from 300 Å for $x = 2/3$ to ≈ 100 Å for $x = 3/4$. The amplitude fluctuation of JTS gives rise to some sort of 'clustering' effect in different regions. Occasionally, small regions without charge-ordering were found, and it is tempting to associate these regions with ferromagnetic metallic clusters. We note that experimental evidences of cluster formation of magnetic polarons have been reported recently in the manganites with $x = 1/3$, which exhibit colossal magnetoresistance effect¹⁷.

Pairing of JTS is an interesting new phenomenon in the charge-ordered state of manganites. Conductivity for the paired JTS is expected to be anisotropic. The new charge-ordering pattern requires new examinations of the spin structure proposed previously^{11,18}. Pairing of JTS appears to be a universal phenomenon in three-dimensional perovskite manganites with a charge-ordered ground state. We have also observed similar pairing of JTS in (Bi,Ca)MnO₃ and (Pr,Ca)MnO₃. Paired JTS are related to charge-ordering in manganites as Cooper pairs are related to superconductivity. Collapse of charge ordering under magnetic field¹⁹, pressure²⁰, and X-ray irradiation²¹ may be due to breaking-up of the paired JTS. The possibility that pair formation of JTS may affect transport and magnetic properties should be carefully considered. The distinct contraction of JTS pairs could be used to identify small residual paired JTS clusters which might form in the lower-doping metallic regime where colossal magnetoresistance occurs. The presence of small JTS clusters in the ferromagnetic-metallic phase could have significant implications for the physics or chemistry of colossal magnetoresistance, which is generally assumed to be a phenomenon of a homogeneous system. □

Received 11 September; accepted 29 December 1997.

- Tranquada, J. M., Sternlieb, B. J., Axe, J. D., Nakamura, Y. & Uchida, S. Evidence for stripe correlations of spins and holes in copper oxide superconductors. *Nature* **375**, 561–563 (1995).
- Wells, B. O. et al. Incommensurate spin fluctuations in high-temperature superconductors. *Science* **277**, 1067–1071 (1997).
- Salkola, M. I., Emery, V. J. & Kivelson, S. A. Implications of charge ordering for single-particle properties of high- T_c superconductors. *Phys. Rev. Lett.* **77**, 155–158 (1996).
- Zaanen, J. & Gunnarsson, O. Charged magnetic domain line and the magnetism of high- T_c oxides. *Phys. Rev. B* **40**, 7391–7394 (1989).
- Dai, P., Mook, H. A. & Dogan, F. Pseudogap and incommensurate magnetic fluctuations in $YBa_2Cu_3O_{6.6}$. *Phys. Rev. Lett.* (submitted).
- Chen, C. H., Cheong, S.-W. & Hwang, H. Y. Charge-ordered stripes in $La_{1-x}Ca_xMnO_3$ with $x > 0.5$. *J. Appl. Phys.* **81**, 4326–4330 (1997).
- Jin, S. et al. Thousandfold change in resistivity in magnetoresistive La-Ca-Mn-O films. *Science* **264**, 413–415 (1994).
- Kuwahara, H., Tomioka, Y., Asamitsu, A., Moritomo, Y. & Tokura, Y. A first-order phase transition induced by a magnetic field. *Science* **270**, 961–963 (1995).
- Ramirez, A. P. et al. Thermodynamic and electron diffraction signatures of charge and spin ordering. *Phys. Rev. Lett.* **76**, 3188–3191 (1996).
- Chen, C. H. & Cheong, S.-W. Commensurate to incommensurate charge ordering and its real-space images in $La_{0.5}Ca_{0.5}MnO_3$. *Phys. Rev. Lett.* **76**, 4042–4045 (1996).
- Goodenough, J. B. Theory of the role of covalence in the perovskite-type manganites $[La,M(II)]MnO_3$. *Phys. Rev.* **100**, 564–573 (1955).
- Radaelli, P. G., Cox, D. E., Marezi, M. & Cheong, S.-W. Charge, orbital, and magnetic ordering in $La_{0.5}Ca_{0.5}MnO_3$. *Phys. Rev. B* **55**, 3015–3023 (1997).
- Ritter, C. et al. Influence of the oxygen content on the structural, magneto-transport, and magnetic properties of $LaMnO_{3+\delta}$. *Phys. Rev. B* **56**, 8902–8911 (1997).
- Egami, T., Louca, D. & McQueeney, R. J. Nonuniform metallic states in manganites and cuprates. *J. Supercond.* **10**, 323–327 (1997).
- Tyson, T. A. et al. Evidence for a local lattice distortion in Ca-doped $LaMnO_3$. *Phys. Rev. B* **53**, 13985–13988 (1996).
- McMillan, W. L. Theory of discommensurations and the commensurate-incommensurate charge-density-wave phase transition. *Phys. Rev. B* **14**, 1496–1502 (1976).
- De Teresa, J. M. et al. Evidence for magnetic polarons in the magnetoresistive perovskites. *Nature* **386**, 256–259 (1997).
- Wollan, E. O. & Koehler, W. C. Neutron diffraction study of the magnetic properties of the series of perovskite-type compounds $[(1-x)La_xCa]MnO_3$. *Phys. Rev.* **100**, 545–562 (1955).
- Tomioka, Y., Asamitsu, A., Moritomo, Y., Kuwahara, H. & Tokura, Y. Collapse of a charge-ordered state under a magnetic field in $Pr_{1/2}Sr_{1/2}MnO_3$. *Phys. Rev. Lett.* **74**, 5108–5111 (1995).
- Moritomo, Y., Kuwahara, H., Tomioka, Y. & Tokura, Y. Pressure effects on charge-ordering transitions in perovskite manganites. *Phys. Rev. B* **55**, 7549–7556 (1997).
- Kiryukhin, V. et al. An X-ray-induced insulator-metal transition in a magnetoresistive manganite. *Nature* **386**, 813–815 (1997).

Acknowledgements. S.M. is on leave from the Department of Physics, Tokyo Institute of Technology. We thank B. Batlogg, P. B. Littlewood and B. I. Schraiman for discussions. One of us (S.M.) thanks the Yamada Science Foundation for financial support.

Correspondence and requests for materials should be addressed to C.H.C. (e-mail: chchen@physics.bell-labs.com).

Electrically switchable mirrors and optical components made from liquid-crystal gels

R. A. M. Hikmet & H. Kemperman

Philips Research, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands

In liquid-crystal (LC) display devices, patterned electrodes are used to effect switching of molecular orientation within a pixel element, and thin layers of a material (typically a polymer) at the surfaces of the cell plates induce the liquid-crystal molecules to revert to their original orientation after the electric field is switched off¹. Because of their periodic variation in refractive index, cholesteric LC phases (which have a helical variation in orientation) reflect light at a wavelength determined by the helical pitch^{2,3}, and so can potentially be used in switchable optical devices such as shutters, reflectors and notch- and band-pass filters. But orientation layers are unable to restore the initial orientation in cholesteric phases. They can instead be given an internal 'memory' of their initial orientation by creating an anisotropic polymer network within the system using *in situ* photopolymerization⁴. Here we show that such crosslinked cholesteric gels can be used to produce fast electrically switchable reflectors with narrow- to broad-band widths (the latter having a silvery mirror-like appearance). By photopolymerizing in a patterned manner, we can make image recordings in the gels which become visible on application of an electric field. These patterned gels offer the prospect of making optical components such as lenses and gratings by holographic recording.

A cholesteric liquid-crystal phase is obtained when a nematic phase is doped with chiral molecules. In the cholesteric phase, the long axis of the liquid-crystal molecules (the director $\mathbf{n}$) rotates about a helix. In this phase a band of incident circularly polarized light having the same sense as the cholesteric helix is reflected whereas the band with the opposite sense is transmitted. The upper ($\lambda_{\max}$) and the lower ($\lambda_{\min}$) boundaries of the reflected band are $\lambda_{\max} = p \times n_e$ and $\lambda_{\min} = p \times n_o$ respectively, where p is the cholesteric pitch corresponding to the length over which the director rotates 360°, and n_e and n_o are the extraordinary and the ordinary refractive indices of a uniaxially oriented phase respectively. The pitch is determined by the concentration of the chiral component and decreases with increasing chiral fraction. The reflected bandwidth ($\Delta\lambda$) is given by $\Delta\lambda = \lambda_{\max} - \lambda_{\min} = p \times (n_e - n_o)$.

In conventional LC cells, long-range orientation of LC molecules is induced by orientation layers. The orientation of the molecules is altered by applying an electric field across transparent electrodes present on the cell surfaces underneath the orientation layers. Upon removal of the field, the LC molecules revert back to the initial orientation state under the influence of these orientation layers. When cholesteric LC is switched from a defect-free planar orientation state, it is almost impossible for the system to reorient itself back to the initial state under the influence of the surface orientation layers. Instead, the cholesteric helix becomes oriented in various directions and the cell shows a scattering texture^{2,3}. To obtain fast and successful switching, a memory state needs to be built into such a cholesteric system.

We achieved this by creating a lightly crosslinked network dispersed within the non-reactive LC molecules. This is done by *in situ* polymerization of a liquid-crystal monoacrylate and diacrylate mixture in the presence of non-reactive LC molecules. After obtaining defect-free orientation of the mixture, polymerization of the reactive molecules is induced by ultraviolet irradiation freezing in the cholesteric configuration and orientation by creating

a network containing non-reactive molecules (anisotropic gel). As opposed to the gel systems we studied previously⁴, in these gels the network consists of monoacrylate molecules forming the side-chain polymer that are crosslinked by a very small amount of diacrylates. The network is in strong interaction with the non-reactive LC which can be switched together with the side-chain polymer upon application of an electric field. In these gels the function of the diacrylate molecules that are present at fractions of a per cent is twofold: (1) they form crosslinks and thus provide a system memory; (2) they preserve the polymer structure and its distribution within the system, thus preventing its diffusion. This second function is especially important in producing broad-band reflectors and patterned gels, as described below.

We studied the switching behaviour of the gels by using visible spectroscopy. Figure 1 shows the switching modes obtained for cholesteric gels in which continuous reversible change in the reflection characteristics of the gels can be induced. In the first mode of switching (Fig. 1, top), the reflection band shifts gradually

to low wavelengths with increasing voltage before decreasing in magnitude and shifting back to higher wavelengths and disappearing at higher voltages. This behaviour is associated with the tilting of the cholesteric helix followed by its unwinding. In the second mode (Fig. 1, bottom), the reflection band becomes narrower and decreases in magnitude as only the lower limit of the reflection band keeps its position with increasing voltage. This indicates that the cholesteric structure and the helical pitch remains the same while molecules tilt to become oriented in the direction of the applied field. Even though this so-called conical deformation mode forms the basis for 90° twisted and supertwisted nematic displays, to our knowledge this is the first report of its observation for tight-pitch cholesterics. These two different type of switching were obtained for the monoacrylates shown in the corresponding figures. From theoretical considerations⁵, the two different types of switching are dependent on various material constants and the cell thickness. We have shown that various parameters such as the cell thickness, the pitch and elastic constants can be adjusted to change the switching

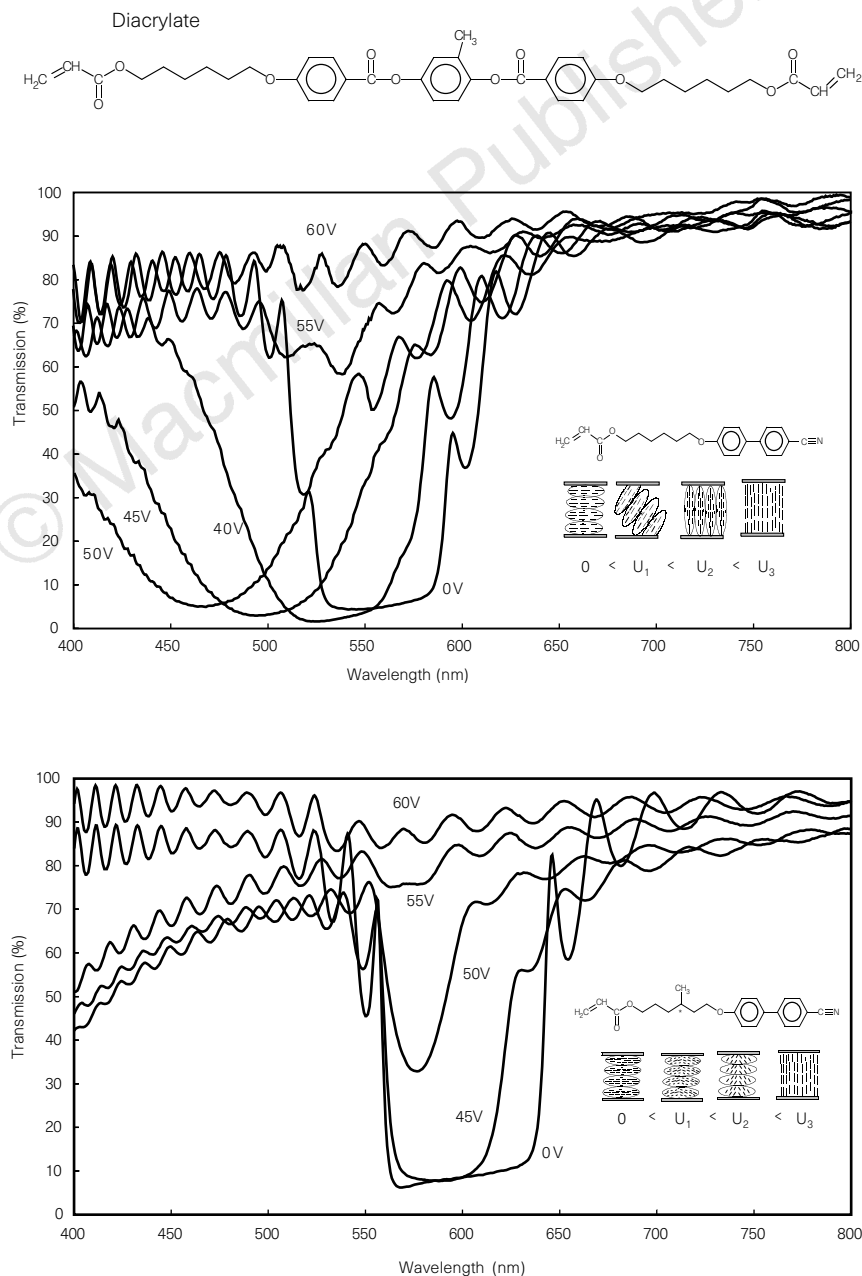


Figure 1 Transmission spectra as a function of applied voltage of cholesteric gels obtained by using two different kinds of monoacrylates and the diacrylate shown.

Drawings show the processes taking place within the gels during switching at various applied voltages of U .

between the modes described above. In both cases, upon removal of the voltage, fast switching (~ 10 ms) of the gels back to the defect free initial orientation state was observed.

As explained above, the width of reflection band of a cholesteric LC is limited to ~ 80 nm in the visible region owing to the birefringence of the materials. In these gels we induced phase separation so as to induce an inhomogeneous distribution of the chiral component within the system, which determines the chole-

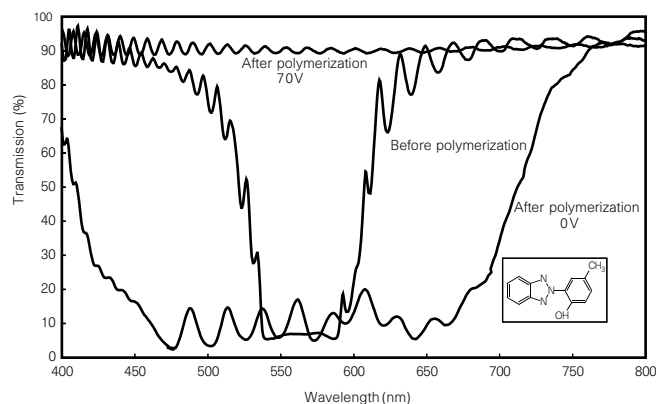


Figure 2 Transmission spectra before and after polymerization of a cholesteric system at zero voltage and the polymerized system at 70 V. The excited-state quencher is also shown.

teric pitch. It is well known that with increasing molecular mass of the polymer its miscibility with monomer decreases, leading to phase separation. In the gels during polymerization, such a phase separation leads to concentration fluctuations corresponding to regions with different pitches. These fluctuations are fixed by the presence of crosslinks and the system remains kinetically stable thereafter. Various parameters influencing the kinetic chain length (molecular mass) of the polymer, such as the initiator concentration and the ultraviolet intensity, are found to influence the width of the reflection band. We found that when compounds that we refer to as excited-state quenchers were added to the monomeric mixtures, there was a further increase in the band width. Such quenchers work by transferring the energy of the excited state of the initiator to themselves, thereby reducing radical formation. They also absorb light, causing an inhomogeneous intensity of light distribution and hence influencing the speed of polymerization across the cell during polymerization⁶. Figure 2 shows the reflection bands of a cholesteric mixture before and after polymerization: it can be seen that upon polymerization, the reflection bands become broader. These broad-band cholesteric gels can be switched reversibly between silver-coloured, reflecting and non-reflecting, transparent states (Fig. 2). A photograph of such a reflector in the reflecting and non-reflecting states observed for circularly polarized light is shown in Fig. 3. Note that the gels were stable to further irradiation with ultraviolet light and did not change their characteristics. The polymer network also showed good thermal stability and even after heating the system well above its clearing temperature, upon cooling the system it reverted back to its initial state.

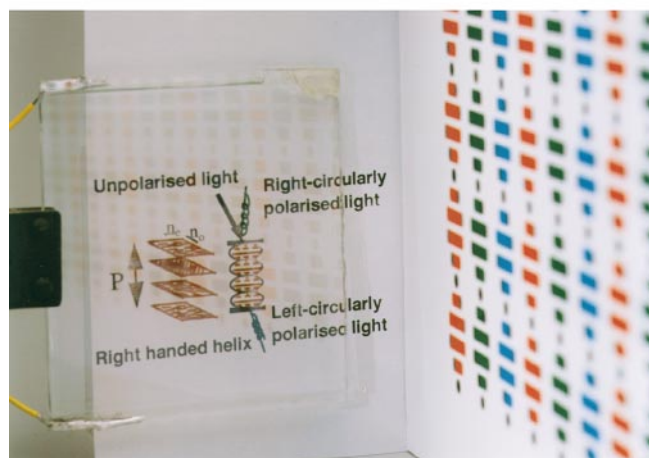
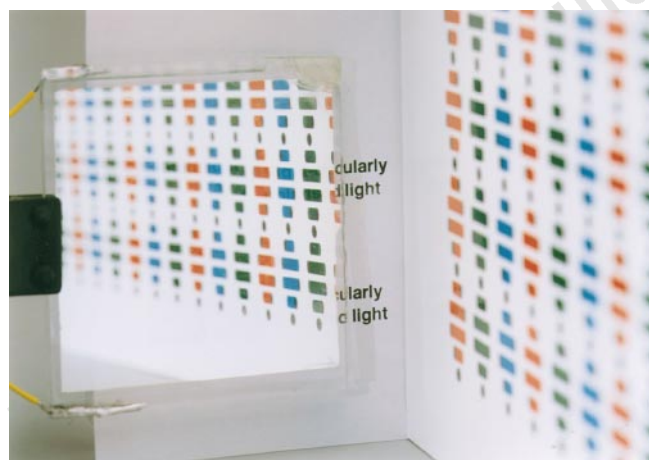


Figure 3 Photographs of an object reflected by a broad-band reflector with the voltage off (top) and on (bottom).

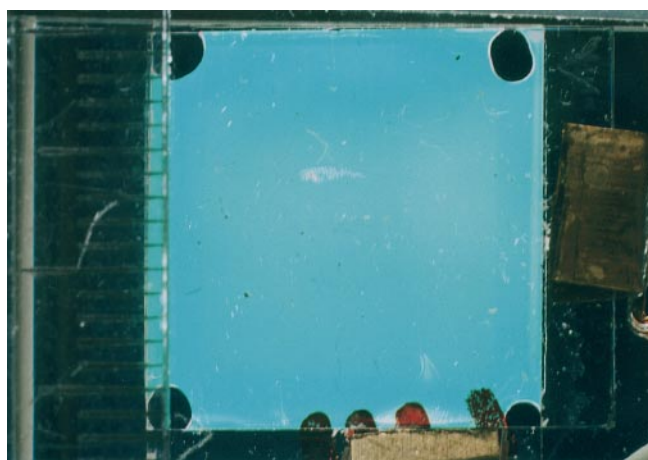


Figure 4 Pattern-wise polymerized gel as observed with the voltage off (top) and on (bottom).

We also investigated the possibility of making recordings in the gels by pattern-wise irradiation of the mixture followed by flood exposure. During the first exposure, polymerization takes place in irradiated areas and diffusion of reactive molecules to these areas occurs owing to formation of a concentration gradient of unreacted species. After flood exposure, the rest of the reactive molecules within the system become polymerized. As a result of diffusion and the nature of the network formed at various steps of polymerization, the threshold voltage for switching in various regions becomes different. In this way, complicated patterns can be recorded into such a gel and made visible upon application of an electric field. We have used this technique in the production of a switchable binary-phase planar Fresnel lens⁷. A cell containing the monomeric mixture was irradiated through a mask containing concentric rings defining the zones of the lens placed above the cell. Subsequently the mask was removed and the cell was exposed to ultraviolet light. As a result, the structure became transferred into the gel in the form of regions with high and low switching voltages. Applying a voltage activated the lens function, as neighbouring regions with high and low refracting indices are formed within the cell. Upon application of an electric field, a beam of light could be focused to a point. In the same way, more complicated recordings could be made. Figure 4 shows a recording made into a gel by illumination through a photo-negative: before application of an electric field no recording is visible, but application of an electric field switches the areas with low threshold voltage, making the image visible. □

Received 17 November 1997; accepted 21 January 1998.

- Gray, G. W. in *Thermotropic Liquid Crystals* 74 (Wiley, Chichester, 1987).
- Gerber, P. R. Voltage induced cholesteric structure transformation in thin layers. *Z. Naturforsch.* **36**, 718–726 (1981).
- Schadt, M. & Gerber, P. Dielectric electro-optical and phase change properties of liquid crystal guest host displays. *Mol. Cryst. Liq. Cryst.* **65**, 241–263 (1981).
- Hikmet, R. A. M. & Lub, J. Anisotropic networks and gels obtained by photopolymerisation in the liquid crystalline state. *Prog. Polym. Sci.* **21**, 1165–1209 (1996).
- Kelker, H. & Hatz, R. in *Handbook of Liquid Crystals* 172 (Verlag Chemie, Weinheim, 1980).
- Broer, D. J., Lub, J. & Mol, G. N. Wide-band reflective polarizers from cholesteric polymer networks with a pitch gradient. *Nature* **378**, 467–469 (1995).
- Patel, J. S. & Rastani, K. Electrically controlled polarisation independent liquid crystal Fresnel lens arrays. *Optics Lett.* **16**, 532–534 (1991).

Acknowledgements. We thank L. Poels for technical assistance and J. Lub for synthesis of the reactive molecules.

Correspondence and requests for materials should be addressed to R.A.M.H. (e-mail: hikmet@natlab.research.philips.com).

Light-driven production of ATP catalysed by F_0F_1 -ATP synthase in an artificial photosynthetic membrane

Gali Steinberg-Yfrach*, Jean-Louis Rigaud†, Edgardo N. Durantini*, Ana L. Moore*, Devens Gust* & Thomas A. Moore*

* Department of Chemistry and Biochemistry and Center for the Study of Early Events in Photosynthesis, Arizona State University, Tempe, Arizona 85287-1604, USA

† Institut Curie, Section de Recherche, UMR-CNRS 168 and LCR-CEA 8, 11 rue Pierre et Marie Curie, 75231 Paris Cedex 05, France

Energy-transducing membranes of living organisms couple spontaneous to non-spontaneous processes through the intermediacy of protonmotive force (p.m.f.)—an imbalance in electrochemical potential of protons across the membrane. In most organisms, p.m.f. is generated by redox reactions that are either photochemically driven, such as those in photosynthetic reaction

centres, or intrinsically spontaneous, such as those of oxidative phosphorylation in mitochondria. Transmembrane proteins (such as the cytochromes and complexes I, III and IV in the electron-transport chain in the inner mitochondrial membrane) couple the redox reactions to proton translocation, thereby conserving a fraction of the redox chemical potential as p.m.f. Many transducer proteins couple p.m.f. to the performance of biochemical work, such as biochemical synthesis and mechanical and transport processes. Recently, an artificial photosynthetic membrane was reported in which a photocyclic process was used to transport protons across a liposomal membrane, resulting in acidification of the liposome's internal volume¹. If significant p.m.f. is generated in this system, then incorporating an appropriate transducer into the liposomal bilayer should make it possible to drive a non-spontaneous chemical process. Here we report the incorporation of F_0F_1 -ATP synthase into liposomes containing the components of the proton-pumping photocycle. Irradiation of this artificial membrane with visible light results in the uncoupler- and inhibitor-sensitive synthesis of adenosine triphosphate (ATP) against an ATP chemical potential of $\sim 12 \text{ kcal mol}^{-1}$, with a quantum yield of more than 7%. This system mimics the process by which photosynthetic bacteria convert light energy into ATP chemical potential.

The artificial photosynthetic membrane is shown schematically in Fig. 1, which illustrates a portion of the liposomal bilayer, the proton-pumping photocycle, and the F -type ATP synthase enzyme—the bulbous structure, top right. The proton pump is driven by vectorial photoinduced electron transfer in a carotene-porphyrin-naphthoquinone molecular triad (**1**, C-P-Q), which generates the species C^+-P-Q^- on excitation by visible light^{1,2}. The

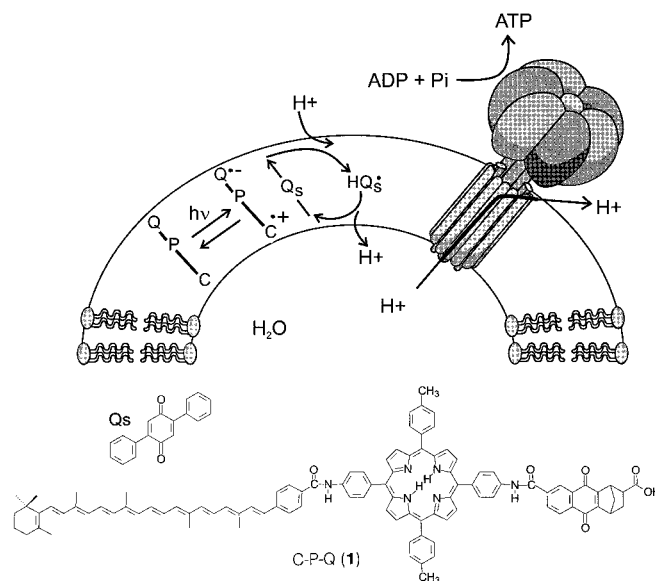


Figure 1 Diagram of a liposome-based artificial photosynthetic membrane. Included in this diagram is the photocycle that pumps protons into the interior of the liposome and the F_0F_1 -ATP synthase enzyme. Triad **1** and Qs were as reported previously¹. For most of the experiments reported here, the triad used consisted of a mixture of isomers in which the carboxylate group was in the 2 and 3 positions. Experimental details are given in the Methods section. Experiments in which a diffusion potential was established before irradiation by adding K_2SO_4 to a final concentration of 0.1 M and valinomycin to the external solution resulted in no change in the rate of ATP synthesis, within experimental error. This is consistent with electrogenic proton import as proposed herein in which the necessary p.m.f. is established by the proton pump. The ATP synthase was activated by addition of $10 \mu\text{M}$ thioredoxin. Oxygen was removed from the cuvette by flow argon. The ATP synthase was activated by addition of $10 \mu\text{M}$ thioredoxin.

intramolecular redox potential represented by the carotenoid radical cation and the naphthoquinone radical anion is coupled to proton translocation by a lipophilic quinone (Qs) which alternates between reduced and oxidized states. Reduction occurs near the external bilayer–water interface when Qs accepts an electron from the naphthoquinone radical anion of **1** to form Qs^{•−}. After protonation of Qs^{•−} near the external aqueous interface, the semiquinone HQs[•], either by diffusion or self-exchange reactions among the Qs molecules, delivers the proton and electron to the site of oxidation potential (C⁺–P–Q) located near the inner membrane surface. Oxidation of HQs[•] near the internal aqueous interface results in proton ejection to the intraliposomal volume and the generation of p.m.f. (These steps only outline the photocycle. Other species could be involved, including a mixture of oxidation states of Qs, provided that the mixture includes protonated species.) On accumulation of sufficient p.m.f., proton flow through the coupling factor with the concomitant formation of ATP from adenosine diphosphate (ADP) and inorganic phosphate (P_i) is thermodynamically allowed.

Proteoliposomes were prepared by incorporation of CF₀F₁-ATP synthase and the proton pump into liposomes. The solutions were illuminated with laser light at a wavelength of 633 nm for a period of time and then assayed for ATP. The synthesis of ATP was measured by the luciferin/luciferase assay³. Figure 2 shows the oxyluciferin luminescence spectrum, recorded as a function of irradiation time, for a sample of proteoliposomes containing the components shown in Fig. 1, including ADP and P_i in the external solution. The increase in luminescence with exposure to actinic light is clear evidence of an increase in ATP concentration with irradiation.

To establish that this increase was due to p.m.f.-linked synthesis of ATP, a number of control experiments were carried out. Samples containing an uncoupler (1 μM carbonyl cyanide *p*-trifluoromethoxyphenylhydrazone (FCCP), which renders the membrane permeable to protons and thereby prevents the build-up of p.m.f.) or an inhibitor (2 μM tentoxin, which is a specific inhibitor of CF₀F₁-ATP synthase) evidenced no ATP production⁴. Other control experiments included deletion of either Qs, ADP, P_i, triad **1** or ATP synthase. In each case the luminescence spectrum after irradiation

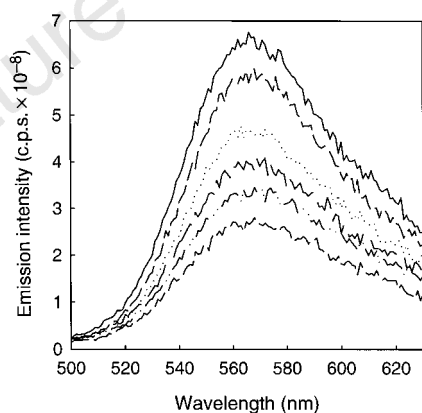


Figure 2 The [ATP]-dependent steady-state luminescence of oxyluciferin³ measured as a function of liposome irradiation times. Shown are data for irradiation times of 0 min (short dashes), 3 min (dash-dot-dot), 6 min (long dashes), 12 min (dots), 35 min (medium dashes) and 66 min (solid line). The irradiation source was a 5 mW beam of 633-nm light from a HeNe laser, and the sample volume was 250 μl (c.p.s., counts per second). The laser beam diameter was expanded to ~10 mm, but no effort was made to provide uniform intensity across the cuvette (as discussed in the text, this level of actinic light saturated the rate of synthesis of ATP). In this experiment, the ATP concentration before irradiation was ~2 × 10^{−7} M (~10^{−3} times the level of ADP), and was due to ATP contamination in the commercially available ADP. Oxygen was removed from the cuvette by flowing argon.

was superimposable on that of the spectrum of an unirradiated control sample, within experimental error. Therefore, ATP synthesis was observed only in the presence of all of the components indicated in Fig. 1 and is ascribed to light-generated p.m.f. coupled to ATP synthesis by CF₀F₁-ATP synthase.

The efficacy of the artificial energy-transducing membrane can be defined by three parameters: the quantum yield of ATP production, the thermodynamic potential of the newly synthesized ATP, and the turnover number of the enzyme. Experiments to measure the turnover number and the thermodynamic factors were carried out under light-saturated conditions, whereas the quantum yield of ATP was measured under light-limited, linear conditions. Linear and saturated conditions were operationally defined by the data shown in Fig. 3, which shows the results of a series of experiments in which the rate of ATP synthesis was measured as a function of actinic light intensity. The linear region is shown on the left-hand side of the graph and the saturated region on the right-hand side.

From the light-limited rate of ATP synthesis shown on the left-hand side of Fig. 3, it is calculated that one ATP molecule is produced per 14 photons incident on the sample. Because of light scattering by the liposome solution, there is uncertainty in measuring the fraction of the actinic light absorbed. Spectroscopic determinations suggest that ~50% of the incident light was absorbed, so the actual quantum yield of ATP synthesis is probably near 0.15. Assuming that four protons are required per ATP synthesized, the yield of the proton pump must be near 0.6. Therefore, both the quantum yield of redox equivalents from photoinduced electron transfer in **1** and the efficiency of proton shuttling by Qs must be quite high. The high yield of redox equivalents from the photoinduced electron transfer step could arise in part from the slowing of energy-wasting charge recombination due to migration of the electrons and holes among the ~1,000 molecules of **1** and the ~500 molecules of Qs in a proteoliposome. This would require some form of aggregation involving **1** and possibly Qs, as rapid intermolecular charge migration requires relatively strong intermolecular electronic interactions. Indeed, the absorption spectrum of **1** in proteoliposomes shows significant broadening of the

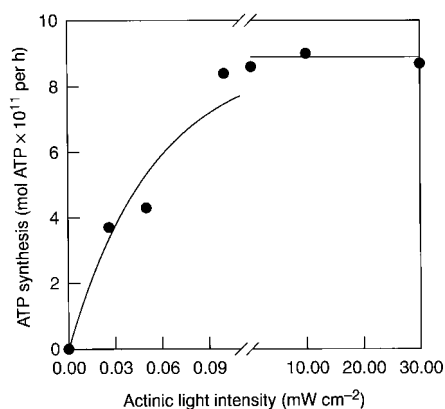


Figure 3 The rate of ATP synthesis as a function of actinic light intensity. Initial conditions are [ATP] = [ADP] = 0.2 mM and [P_i] = 5 mM. Full saturation was reached at actinic light levels of 0.1 mW cm^{−2}. The ATP concentration was determined by the luciferin/luciferase assay shown in Fig. 2. In a typical experiment, aliquots of irradiated solution in which ATP had been formed were withdrawn as a function of irradiation time and added to an assay solution containing luciferin, luciferase and buffer. The luminescence spectrum was measured immediately and the [ATP] determined from a standard curve which was prepared for each experimental run by measuring the steady-state luminescence from preparations having known ATP concentrations. In all cases the steady-state luminescence intensity was constant for at least 20 min.

porphyrin Soret band which is indicative of aggregate formation. Because the quantum yield is high and the aggregated form of **1** is predominant, electron transfer processes between the various oxidation states of Qs and reducing/oxidizing equivalents in aggregates of **1** must drive the proton pump.

Turning to thermodynamic considerations, the chemical potential of ATP synthesized at constant temperature and pressure, under conditions of 0.2 mM ATP, 0.02 mM ADP and 5 mM P_i , is $\sim 12 \text{ kcal mol}^{-1}$ (using concentration as activity in the Gibbs equation). As shown in Fig. 4, synthesis proceeds readily under these conditions. Thus ATP synthesis occurs against a substantial ATP potential. The slightly lower rate of formation of ATP at $[\text{ATP}]/[\text{ADP}] = 10$ compared to that at $[\text{ATP}]/[\text{ADP}] = 1$ could be due to the lower concentration of ADP in the former rather than the higher $[\text{ATP}]/[\text{ADP}]$. That is, the enzyme may not have been saturated with respect to ADP.

The synthesis of ATP at high ATP potential demonstrates that significant p.m.f. is being generated by the proton pumping photocycle. In fact, ATP synthesis against similar ATP potentials has been reported in isolated chloroplast lamellae⁵. Detailed experiments to measure the maximum p.m.f. available from this system are under way.

To determine the turnover number, conditions were selected such that the enzyme was saturated with ADP (0.2 mM) and P_i (5 mM)⁶. The light-saturated rate of ATP synthesis (right side of Fig. 3) was used for the calculation. For maximum activity, non-catalytic ATP binding sites on the F-type ATPases must be occupied^{7–10}. To ensure this, experiments were carried out with 0.2 mM ATP added before irradiation. Under these conditions $3.5 \times 10^{-8} \text{ mol}$ of ATP were synthesized per hour. As there are $\sim 9 \times 10^{11}$ liposomes in the irradiated volume and one enzyme molecule per liposome, the turnover number is $7 \pm 1 \text{ ATP per CF}_0\text{F}_1 \text{ per s}$. This number is higher than has been reported in hybrid systems using natural photosystem I preparations as the proton pump, and is similar to that observed in bacteriorhodopsin/ATPase constructs^{10,11}.

In order to identify factors that control the CF_0F_1 turnover number in this system, pH-jump-induced ATP synthesis in proteoliposome preparations containing all of the components of the proton pump was compared to that in proteoliposomes prepared without **1** and Qs. Both types of liposomes gave turnover numbers of $\sim 100 \text{ s}^{-1}$. This level of activity under comparable pH-jump conditions has been reported for similar proteoliposomes, which indicates that neither **1** nor Qs inhibits the enzyme¹⁰. Therefore, because the pH-jump experiment illustrates that the enzyme is

capable of higher turnover, and the synthesis of ATP against high ATP potential illustrates that the p.m.f. is high, the light saturation of ATP synthesis ($\sim 0.1 \text{ mW cm}^{-2}$; right side of Fig. 3) at a relatively low turnover number probably arises from kinetic constraints in the photocyclic proton pump. The dynamics of the photocyclic proton pump have not been investigated, but saturation effects could be due to electron-hole annihilation in the aggregated arrays of **1** in the membrane or to diffusion-limited transport processes.

We note that the photocyclic system operates efficiently over a timescale of hours. Figure 4 shows ATP synthesis as a function of irradiation time under conditions of maximum enzyme turnover. A single ATP synthase turns over more than 25,000 times per hour. Also shown are the data for control experiments carried out at high initial $[\text{ATP}]$. As was the case when ATP was not added before irradiation, ATP synthesis is uncoupler- and inhibitor-sensitive and only observed when all of the components indicated in Fig. 1 are present.

Based on the ATP potential, the quantum yield of formation, and the energy of 633-nm photons we estimate that up to 4% of the absorbed light energy is conserved in this system. The steps in the energy transduction process in the artificial membrane mimic those by which photosynthetic bacteria convert light energy into redox potential, p.m.f. and ultimately chemical energy available as ATP. The assembly of an artificial photosynthetic membrane that demonstrates net energy conservation opens the door to the design of systems in which energy-requiring biological and biomimetic processes in cellular-like constructs could be driven from a photocyclic energy source. □

Methods

Proteoliposomes containing CF_0F_1 -ATPase isolated from spinach chloroplasts were prepared by adding liposomes to detergent-solubilized protein and slowly removing the detergent with Biobeads¹². This procedure results in the incorporation of an average of one enzyme molecule per liposome, with the F_1 subunit having the indicated preferred orientation extending into the bulk aqueous phase¹². The liposomes were prepared from a mixture of egg phosphatidylcholine and egg phosphatidic acid (10:1 by weight) plus cholesterol (20 mol%)¹². For experiments where ATP synthesis was observed, Qs was added to the phospholipid mixture before liposome formation. Triad **1** was added to the proteoliposomes by injection of $10 \mu\text{l}$ of a 3.8 mM solution of **1** in tetrahydrofuran into $250 \mu\text{l}$ of a solution containing proteoliposomes followed by column chromatography (Sephadex G-100) to remove excess triad. Approximately $800 \mu\text{l}$ of solution was collected from the column. Analysis of light-scattering experiments on similar liposome preparations that contained chlorophyll rather than Qs or **1** yielded a most probable diameter of 150 nm (ref. 10). Total phosphate analysis of $250 \mu\text{l}$ of a proteoliposome preparation containing triad **1** and Qs prepared for a typical photochemical experiment yielded $\sim 9 \times 10^{11}$ liposomes (assuming 150-nm diameter liposomes and an area of $58 \text{ \AA}^2$ for phosphatidylcholine and phosphatidic acid and $38 \text{ \AA}^2$ for cholesterol)¹³. Extraction and spectrophotometric assay yielded $\sim 1 \times 10^3$ molecules of **1** per liposome, which is many more than the ~ 40 per liposome previously found in liposomes formed from other phospholipids and lacking the protein component¹. Based on the amount of Qs added before forming the liposomes, there were no more than 500 Qs molecules per liposome. The bulk aqueous phase contained 16 mM Na_2SO_4 , 2.5 mM MgSO_4 and was buffered by 16 mM tricine and 5 mM KH_2PO_4 to a pH of 8. The intraliposomal volume contained 2 mM tricine and 40 mM Na_2SO_4 with the pH adjusted to 8 before liposome formation.

Received 13 November 1997; accepted 16 February 1998.

- Steinberg-Yfrach, G. *et al.* Conversion of light energy to proton potential in liposomes by artificial photosynthetic reactions centres. *Nature* **385**, 239–241 (1997).
- Gust, D., Moore, T. A. & Moore, A. L. Molecular mimicry of photosynthetic energy and electron transfer. *Acc. Chem. Res.* **26**, 198–205 (1993).
- Schmidt, G. & Gräber, P. The rate of ATP synthesis by reconstituted CF_0F_1 liposomes. *Biochim. Biophys. Acta* **808**, 46–51 (1985).
- Sigalat, C., Pitard, B. & Haraux, F. Proton coupling is preserved in membrane-bound chloroplast ATPase activated by high concentrations of tentoxin. *FEBS Lett.* **368**, 253–256 (1995).
- Hangarter, R. P. & Good, N. E. Energy thresholds for ATP synthesis in chloroplasts. *Biochim. Biophys. Acta* **681**, 397–404 (1982).

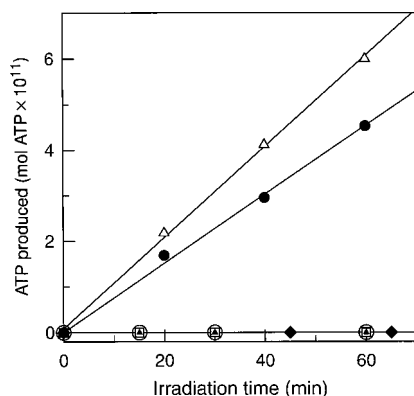


Figure 4 The synthesis of ATP as a function of irradiation time at different ratios of $[\text{ATP}]/[\text{ADP}]$. Open triangles, $[\text{ATP}] = [\text{ADP}] = 0.2 \text{ mM}$ and $[P_i] = 5 \text{ mM}$; filled circles, $[\text{ATP}] = 0.2 \text{ mM}$, $[\text{ADP}] = 0.02 \text{ mM}$ and $[P_i] = 5 \text{ mM}$. Also shown are results for control experiments showing that ATP was not synthesized; open circles, $1 \mu\text{M}$ FCCP; filled triangles, $2 \mu\text{M}$ tentoxin; filled diamonds, $[\text{Qs}] = 0$; open squares, $[\text{ADP}] = 0$.

6. Richard, P. & Gräber, P. Kinetics of ATP synthesis catalyzed by the H^+ -ATPase from chloroplasts (CF_0F_1) reconstituted into liposomes and coreconstituted with bacteriorhodopsin. *Eur. J. Biochem.* **210**, 287–291 (1992).
7. Richard, P., Pitard, B. & Rigaud, J.-L. ATP synthesis by the F_0F_1 -ATPase from the thermophilic *Bacillus* PS3 co-reconstituted with bacteriorhodopsin into liposomes. *J. Biol. Chem.* **270**, 21571–21578 (1995).
8. Fromme, P. & Gräber, P. Activator/inactivation and uni-site catalysis by the reconstituted ATP-synthase from chloroplasts. *Biochim. Biophys. Acta* **1016**, 29–42 (1990).
9. Girault, G., Berger, G., Galmiche, J.-M. & Andre, F. Characterization of six nucleotide-binding sites on chloroplast coupling factor 1 and one site on its purified β subunit. *J. Biol. Chem.* **263**, 14690–14695 (1988).
10. Cladera, J. *et al.* Functional reconstitution of photosystem I reaction center from cyanobacterium *Synechocystis* sp. PCC6803 into liposomes using a new reconstitution procedure. *J. Bioenerg. Biomembr.* **28**, 503–515 (1996).
11. Pitard, B., Richard, P., Duñach, M. & Rigaud, J.-L. ATP synthesis by the F_0F_1 ATP synthase from thermophilic *Bacillus* PS3 reconstituted into liposomes with bacteriorhodopsin. *Eur. J. Biochem.* **235**, 779–788 (1996).
12. Richard, P., Rigaud, J.-L. & Gräber, P. Reconstitution of CF_0F_1 into liposomes using a new reconstitution procedure. *Eur. J. Biochem.* **193**, 921–925 (1990).
13. Petitou, M., Tuy, F. & Rosenfeld, C. A simplified procedure for organic phosphorus determination from phospholipids. *Anal. Biochem.* **91**, 350–353 (1978).

Acknowledgements. We thank F. Haraux for the gift of purified CF_0F_1 , and both F. Haraux and W. Frasch for discussions. This work was supported by the US DOE.

Correspondence and requests for materials should be addressed to T.A.M. (e-mail: tmoore@asu.edu).

Ocean margins as a significant source of organic matter to the deep open ocean

James E. Bauer* & Ellen R. M. Druffel†

* School of Marine Science, College of William and Mary, Gloucester Point, Virginia 23062, USA

† Department of Earth System Science, University of California, Irvine, California 92697, USA

Continental shelves and slopes comprise less than 20% of the world ocean area, yet they are proposed to be quantitatively important sources of the organic matter that fuels respiration in the open ocean's interior^{1,2}. At least certain regions of the coastal ocean produce more organic carbon than they respire³, suggesting that some fraction of this non-respired, unburied organic carbon is available for export from the coastal to the open ocean⁴. Previous studies of carbon fluxes in ocean margins^{1,5,6} have not considered the potential roles of dissolved organic carbon (DOC) and suspended particulate organic carbon (POC_{susp}), even though both pools are quantitatively far larger than sinking POC. Here we report natural radiocarbon (^{14}C) abundance measurements that reveal continental slope and rise waters to contain both DOC and POC_{susp} that are concurrently older and in higher concentrations than DOC and POC_{susp} from the adjacent North Atlantic and North Pacific central gyres. Mass-balance calculations suggest that DOC and POC_{susp} inputs from ocean margins to the open ocean interior may be more than an order of magnitude greater than inputs of recently produced organic carbon derived from the surface ocean. Inputs from ocean margins may thus be one of the factors contributing to the old apparent age of organic carbon observed in the deep North Atlantic and Pacific central gyres^{7–9}.

The radioisotopic form of carbon, ^{14}C (half-life, 5,730 yr), can be used as an indicator of the average age of bulk marine organic carbon pools such as DOC and POC. In addition to natural cosmogenically produced ^{14}C , the increase in the global inventory of ^{14}C as a result of nuclear weapons testing during the late 1950s and early 1960s also allows us to use bomb-produced ^{14}C to constrain the age of more recently formed (over the past ~40 years) organic carbon pools¹⁰. The $\Delta^{14}C$ (defined as the deviation in parts per thousand (‰) from the ^{14}C activity of nineteenth century wood) values of natural marine organic carbon have been found to range from as low as around –525‰ (^{14}C age of ~6,000 yr BP) for

deep ocean DOC⁸ to as high as about +140‰ for suspended POC samples containing bomb ^{14}C in the mid- to late 1980s (ref. 9).

Samples for ^{14}C isotope analysis of DOC and POC_{susp} were collected from the western North Atlantic (WNA) and eastern North Pacific (ENP) continental margins. In April 1994, samples were collected from 4 depths at each of three WNA sites located over the continental slope in the Middle Atlantic Bight region between eastern Long Island, New York, and Cape Hatteras, North Carolina. The Bight is characterized by a permanent thermohaline front between continental shelf and slope waters and a net southwestward flow of slope water, of which ~50% is entrained across the front into slope waters and the remainder is advected offshore near Cape Hatteras⁵. The depth of the WNA mid-slope sites ranged from ~1,100–1,500 m. Samples were also collected from the eastern North Pacific from 1991 to 1993 at a time-series site located at the base of the continental rise (~4,100 m depth) at 34° 50' N, 123° 00' W^{11,12}, and previously in 1985 at a site in the Santa Monica basin of the California continental borderland¹³. The ENP site is located ~220 km off the coast of central California within the California current system and is influenced by spring–summer maxima in primary productivity and sinking POC fluxes as a result of seasonal upwelling¹⁴.

The $\Delta^{14}C$ values of DOC from shallow surface waters (5 m depth) of the WNA continental slope were highly variable, ranging over ~200‰ (Fig. 1a). In mid-depth slope waters (~300 m and 750 m; Fig. 1a), $\Delta^{14}C$ -DOC values were significantly lower (that is, more negative in $\Delta^{14}C$) by 75 to 150‰ relative to DOC from similar depths in the central north Atlantic gyre (–276 to –260‰ in the Sargasso Sea (SS)^{8,9}. By 1,000 m depth, the $\Delta^{14}C$ -DOC profiles of WNA slope water and those of SS water converged towards similar values. These data indicate the presence of ^{14}C -depleted DOC at mesopelagic depths in WNA slope waters. On the basis of its $\delta^{13}C$ values ($\delta^{13}C$ range: –21.3 to –22.4‰; J.E.B., unpublished data), this DOC appears to be predominantly marine in origin (fully marine DOC has $\delta^{13}C$ values ranging from about –22 to –18‰).

The $\Delta^{14}C$ values of POC_{susp} from the WNA ranged from –190 to +80‰ and were significantly lower than $\Delta^{14}C$ of POC_{susp} from the SS⁹ (Fig. 1b). $\Delta^{14}C$ values more positive than about –70 to –40‰ are considered to be post-bomb (that is, later than early 1950s to early 1960s thermonuclear weapons testing) because the pre-bomb $\Delta^{14}C$ of the temperate and tropical surface ocean was in this range¹⁵. The $\delta^{13}C$ values in POC_{susp} (range: –22.9 to –24.9‰; J.E.B., unpublished data) suggest that terrestrial carbon (with an assumed average $\delta^{13}C \approx -27$ ‰) may have contributed slightly more to POC_{susp} than to DOC in WNA slope waters. The similar offsets in $\Delta^{14}C$ for both DOC and POC_{susp} between the WNA and the SS (Fig. 1a, b) suggest that the same or related mechanisms may control inputs of ^{14}C -depleted DOC and POC_{susp} to WNA slope waters.

Similar to the WNA, profiles from the ENP also indicate a net depletion in ^{14}C (that is, more negative $\Delta^{14}C$ values) of both DOC and POC_{susp} relative to the central North Pacific (CNP) gyre^{7,9} (Fig. 1c, d). The lowest $\Delta^{14}C$ -DOC values in the ENP occurred, also like the WNA, at shallow to intermediate depths (0–700 m), and $\Delta^{14}C$ -DOC values in the ENP were lower than in the CNP at all depths samples (Fig. 1c). The $\Delta^{14}C$ -DOC values in Santa Monica basin were also, with the exception of the single 850-m sample, lower than values in the CNP (Fig. 1c). Significantly lower $\Delta^{14}C$ values of POC_{susp} were likewise observed at all depths in both the ENP and Santa Monica Basin¹³ relative to the CNP gyre (Fig. 1d). The average difference in $\Delta^{14}C$ -DOC between WNA and SS waters was greater than the corresponding margin-central gyre difference in the North Pacific (compare Fig 1a, c); the margin-central gyre offset in $\Delta^{14}C$ of POC_{susp} was also greater overall in the North Atlantic (compare Fig. 1b, d). The corresponding $\delta^{13}C$ values for DOC (range: –20.5 to –21.7 (ref. 12)) and POC_{susp} (range: –20.0 to –22.9‰ (ref. 11)) in ENP waters were greater (more positive) as a whole than for DOC and POC_{susp} from the WNA.

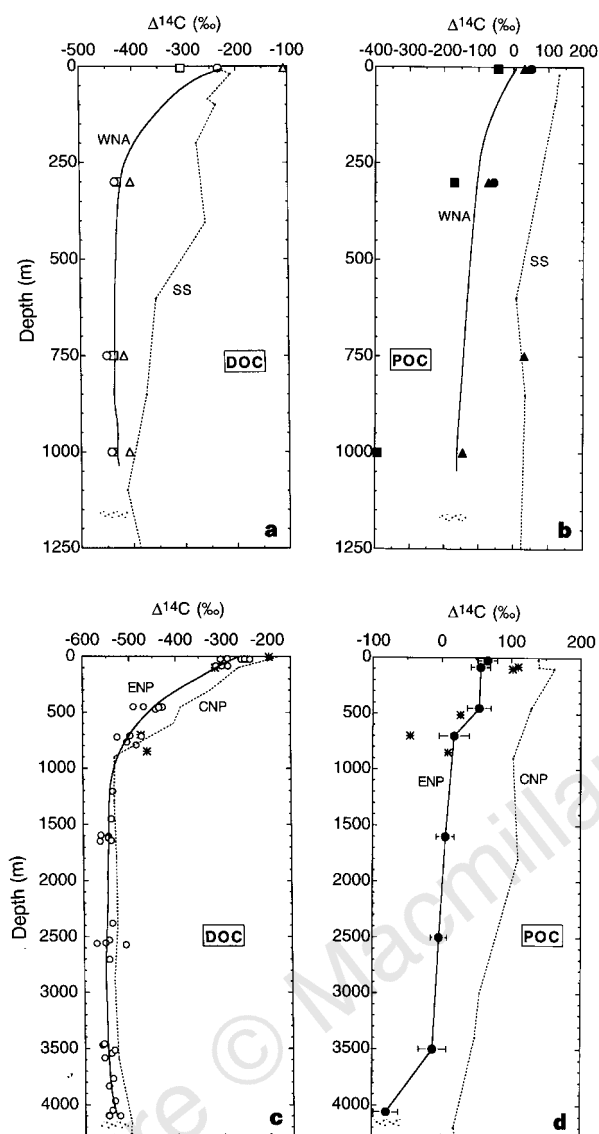


Figure 1 $\Delta^{14}\text{C}$ values of DOC and POC in North Atlantic and North Pacific waters. **a**, Values of $\Delta^{14}\text{C}$ (deviation in parts per thousand from the ^{14}C activity of nineteenth century wood) of DOC (open symbols with solid line-of-best-fit added), and **b**, POC_{susp} (solid symbols with solid line-of-best-fit added) from 3 stations between $35^{\circ}39.06' \text{N}$, $74^{\circ}36.78' \text{W}$ and $39^{\circ}37.72' \text{N}$, $71^{\circ}37.78' \text{W}$ (circles, northern slope station; triangles, central slope station; squares, southern slope station) along the Middle Atlantic Bight continental slope in the western North Atlantic (WNA). Also shown are $\Delta^{14}\text{C}$ -DOC and $\Delta^{14}\text{C}$ - POC_{susp} profiles determined previously for the Sargasso Sea (SS, dashed lines)^{8,9}. DOC and POC_{susp} collection and sample-processing techniques are given in refs 9 and 12. Briefly, DOC samples (650 ml of $0.7\text{-}\mu\text{m}$ -filtered sea water) were oxidized to CO_2 by high-energy (2,400 W) ultraviolet irradiation for 2 h. POC_{susp} samples were collected from up to several thousand litres of sea water using *in situ* pumps with $0.7\text{-}\mu\text{m}$ quartz-fibre filters followed by sealed-tube combustion of sample filters to produce CO_2 . Procedures for conversion of sample CO_2 to graphite and subsequent ^{14}C analysis by accelerator mass spectrometry (AMS) are presented in ref. 30. All ^{14}C data are corrected for sample $\delta^{13}\text{C}$ (data not shown) and are reported using the conventions of Stuiver and Pollach³¹. Errors associated with $\Delta^{14}\text{C}$ AMS analysis are smaller than the symbols and averaged $\pm 6\%$. Errors due to blank corrections for $\Delta^{14}\text{C}$ - POC_{susp} samples from the WNA only (filled symbols) are as follows: 5-m samples, $\leq \pm 11\%$; 300–1,000-m samples, $\pm 40\text{--}60\%$, with the exception of one 300-m sample (-171%), which had an error of -80% . All DOC samples from the WNA slope were significantly depleted in $\Delta^{14}\text{C}$ relative to the SS $\Delta^{14}\text{C}$ -DOC profile, with the exception of the single 1,000-m sample ($\Delta^{14}\text{C} = -408\%$) which has not significantly different, and the single 5-m sample ($\Delta^{14}\text{C} = -107\%$)

Concentrations of DOC were greatest in shallow (0 to $\sim 100 \text{ m}$) WNA (J.E.B., unpublished data) and ENP¹² slope and rise waters and decreased with increasing depth. With the exception of $\sim 300 \text{ m}$ depth in the WNA and 0 to 700 m depth in the ENP, DOC in slope and rise waters exceeded by $4\text{--}9 \mu\text{M}$ those concentrations measured previously in North Atlantic and Pacific central gyre waters^{7–9} (Table 1). Suspended POC concentrations were up to an order of magnitude higher in surface waters of the slope and rise than in surface waters of both the SS (J.E.B., unpublished data; ref. 9) and CNP⁹. At depths $\geq 500 \text{ m}$, POC_{susp} concentrations in slope and rise waters were in all cases $\sim 2\text{--}3$ times ($2\text{--}3 \mu\text{g C l}^{-1}$) greater than in the deep SS and CNP, where POC_{susp} concentrations were $\sim 0.6\text{--}1 \mu\text{g C l}^{-1}$ (Table 1).

The elevated DOC and POC_{susp} concentrations in slope (in the WNA) and rise (in the ENP) waters, together with lower $\Delta^{14}\text{C}$ values in both pools, indicate that organic matter older than that in the North Atlantic and Pacific central gyres is present in ocean margins and that it is potentially available for export to the open ocean. The origin(s) of this old, ^{14}C -depleted carbon to continental slope and rise waters is (are) not known, but several possibilities may be considered. In the WNA, the reintroduction to the water column of old sedimentary organic carbon (both as DOC and POC_{susp}) from weathered shelf and upper slope sediments^{16,17} may contribute to the highly ^{14}C -depleted ($\Delta^{14}\text{C}$ as low as about -700%) colloidal and dissolved organic carbon observed in near-bottom waters in the Middle Atlantic Bight¹⁸ as well as to the ^{14}C -depleted POC_{susp} (Fig. 1b). The WNA may also at times be influenced by upwelled Antarctic Intermediate Water¹⁹ which could contain a component of older, higher-concentration DOC. However, this mechanism is unlikely to account for the elevated concentrations of ^{14}C -depleted POC_{susp} also found in the WNA (Fig. 1b). Finally, the presence of trace amounts of hydrocarbons (with $\Delta^{14}\text{C}$ of about $1,000\%$) from submarine seeps off California²⁰ could impart lower-than-average $\Delta^{14}\text{C}$ values to the DOC pool (but less likely so to the POC_{susp} pool) of the ENP compared with areas remote from seepage (such as the CNP). Thus, the older DOC and POC_{susp} in these two ocean margin systems may arise from multiple system-specific sources or from a common source such as weathered shelf and slope sediments.

The DOC of surface open ocean waters can be shown conceptually to consist of a mixture of both old, conservative material that has aged in the deep ocean and of young labile material recently

which was significantly greater. All POC_{susp} samples from the WNA slope were significantly depleted in $\Delta^{14}\text{C}$ relative to the Sargasso Sea $\Delta^{14}\text{C}$ - POC_{susp} profile, with the exception of the single 750-m sample, which was not significantly different. **c**, Values of $\Delta^{14}\text{C}$ of DOC (open circles with solid line-of-best-fit added), and **d**, POC_{susp} (filled circles) from the eastern North Pacific (ENP) including a continental rise site off central California at $34^{\circ}50' \text{N}$, $123^{\circ}00' \text{W}$ and Santa Monica basin (star symbols) in the southern California continental borderland¹³. Symbols for $\Delta^{14}\text{C}$ of DOC in the ENP (open circles) represent individual samples measured on six separate occasions between July 1991 and July 1993. Symbols for $\Delta^{14}\text{C}$ of POC_{susp} in the ENP (solid circles) represent mean values ($\pm 1\sigma$) of four profiles measured between February 1992 and July 1993 (ref. 11). Also shown are profiles of $\Delta^{14}\text{C}$ -DOC and $\Delta^{14}\text{C}$ - POC_{susp} (both dashed lines) determined previously for the central North Pacific (CNP)^{7,9}. Errors associated with $\Delta^{14}\text{C}$ AMS analyses of both DOC and POC_{susp} are smaller than symbols and averaged $\pm 6\%$ ($\pm 1\sigma$). For $\Delta^{14}\text{C}$ -DOC in the CNP, single profiles measured in 1985 and 1987 showed minimal differences from each other⁹; the average standard deviation of these $\Delta^{14}\text{C}$ -DOC profiles in the upper 1,000 m was 12% , equivalent to the 2σ measurement error. All DOC and POC_{susp} samples from both eastern North Pacific sites were significantly depleted in $\Delta^{14}\text{C}$ relative to the corresponding central N. Pacific profiles, with the exception of the 900-m Santa Monica basin DOC sample which was not significantly greater. Note the differences in $\Delta^{14}\text{C}$ scales between all panels and in depth scales between WNA (in **a** and **b**) and ENP (in **c** and **d**) profiles. Stippled area under WNA and ENP profiles indicates approximate depth of the sea floor at these sites.

produced in the photic zone^{9,21,22}. Using a mass-balance approach, it has been calculated⁹ that surface seawater $\Delta^{14}\text{C}$ -DOC values for the central North Pacific and North Atlantic Oceans are -155‰ and -214‰ , respectively, which agree well with measured values of -153‰ and -210‰ , respectively. A similar calculation is made here for surface (5 m) slope water of the central WNA, which has the highest $\Delta^{14}\text{C}$ -DOC value (-107‰ ; Fig. 1a) of all WNA slope waters examined. The background DOC from deeper (300–1,000 m) waters of the central WNA slope has a mean concentration of $52\text{ }\mu\text{M}$ (J.E.B., unpublished data) and a mean $\Delta^{14}\text{C}$ of -421‰ (Fig. 1a). The DOC concentration of surface slope water is $\sim 91\text{ }\mu\text{M}$ (J.E.B., unpublished data), giving a calculated $\Delta^{14}\text{C}$ of the 'excess' surface DOC component (with a concentration of $91\text{ }\mu\text{M}$ minus $52\text{ }\mu\text{M}$, or $39\text{ }\mu\text{M}$) of $+311\text{‰}$. This $\Delta^{14}\text{C}$ value is high and similar to previous measurements of humic substances in Amazon river water made in 1984 ($\Delta^{14}\text{C} = +230\text{‰}$)²³ and reflects the input of post-bomb, terrestrial carbon to shallow central WNA slope waters which are influenced by inputs from rivers and estuaries.

The same calculation when applied to southern WNA waters yields a different conclusion. Using a deep, background mean DOC concentration of $51\text{ }\mu\text{M}$ (J.E.B., unpublished data) and a $\Delta^{14}\text{C}$ -DOC value of -435‰ (Fig. 1a) and a DOC concentration and $\Delta^{14}\text{C}$ -DOC value in shallow slope waters of $83\text{ }\mu\text{M}$ (J.E.B., unpublished data) and -306‰ (Fig. 1a), respectively, then the $\Delta^{14}\text{C}$ of the 'excess' surface DOC component (with a concentration of $83\text{ }\mu\text{M}$ minus $51\text{ }\mu\text{M}$, or $32\text{ }\mu\text{M}$) is calculated to be -100‰ . This $\Delta^{14}\text{C}$ -depleted component of excess DOC added to southern WNA surface slope water (about 400‰ lower than the 'excess' component in central WNA slope surface waters) is reflective of an older source of carbon to the surface DOC pool, perhaps originating from shelf and slope porewaters and sediments^{16,17,24} that are advected seaward. Similar calculations for POC_{susp} using this simple model are not justified because of the ~ 10 -fold greater concentrations of POC_{susp} in surface compared with deeper waters. For DOC, however, it is clear that sources having highly disparate $\Delta^{14}\text{C}$ may contribute to the surface ocean pool, leading to the high degree of variability observed in the WNA.

The existence of positive concentration gradients between the margins and deep open ocean and between the surface and deep open oceans indicates that both margins and the surface ocean may serve as sources of DOC and POC_{susp} to the deep central gyres (Table 1). We can estimate by ^{14}C mass balance the relative potential

contributions of each of these sources to the deep North Atlantic and Pacific using the following simplifying assumptions: (1) the deep central North Atlantic and Pacific are in steady state with respect to $\Delta^{14}\text{C}$ values and concentrations of DOC and POC_{susp} (refs 7–9); (2) the two dominant sources of DOC and POC_{susp} to the deep central gyres are lateral inputs of ^{14}C -depleted material derived from the margins and vertical inputs of 'modern', ^{14}C -enriched material derived from surface ocean production¹¹; and (3) the margin-to-deep open ocean and surface-to-deep open ocean gradients observed in these studies are representative of the North Atlantic and Pacific as a whole. We find that in order to maintain the observed average DOC $\Delta^{14}\text{C}$ values in the deep central gyres, the input of DOC from the margins is calculated to be as much as 25–100 times that of modern, surface ocean-derived carbon; for POC_{susp} , the contribution from margins is smaller than that for DOC but still 5–19 times greater than the contribution of material from the surface (Table 1). These estimates of margin and surface contributions to the deep open ocean have two principal implications: (1) inputs of 'aged' organic carbon from the margins to the deep open ocean may surpass inputs derived from recent surface ocean production, and (2) in view of the much larger surface-to-deep than margin-to-deep concentration gradients, most young, surface-derived material must be degraded, allowing a smaller but more highly refractory margin component to contribute proportionally more to the deep central gyres.

Lateral transport of organic matter from margins to pelagic and abyssal environments has been invoked previously to help explain carbon and oxygen imbalances in the deep ocean^{14,25}. Transport of ^{14}C -depleted DOC and even POC_{susp} from ocean margins to the central gyres may be facilitated by isopycnal (that is, lateral) eddy diffusion, which can be 10^3 – 10^7 times greater than vertical eddy diffusive transport²⁶. Although vertical transport of recently produced surface material to the central gyres may also be enhanced by such processes as seasonal thermocline breakdown²⁷ and rapidly sinking organic particles^{9,11,14}, we would expect this younger organic carbon to be relatively more susceptible to microbial remineralization^{21,22,28} than older margin-derived material. Thus, although the open ocean undoubtedly receives inputs from both its margins and surface production, the $\Delta^{14}\text{C}$ and concentration profiles of DOC and POC_{susp} in the deep central gyres may be maintained by greater relative inputs from the margins than from recent surface production.

Table 1 Relative contributions of margin and surface ocean organic carbon to the deep central gyres

Dissolved organic carbon						
	Margin component		Modern surface component		Margin fraction*	Surface fraction*
	$\Delta - \Delta^{14}\text{C}_{\text{DOC}}^\dagger$	$\Delta[\text{DOC}]^\ddagger$	$\Delta - \Delta^{14}\text{C}_{\text{DOC}}^\S$	$\Delta[\text{DOC}]^\parallel$		Margin: surface ratio
North Atlantic	-78‰	$9\text{ }\mu\text{M}$	$+520\text{‰}$	$34\text{ }\mu\text{M}$	0.96	0.04
North Pacific	-39‰	$4\text{ }\mu\text{M}^\P$	$+665\text{‰}$	$23\text{ }\mu\text{M}$	0.99	0.01
Suspended particulate organic carbon						
	$\Delta - \Delta^{14}\text{C}_{\text{POC}}^\dagger$	$\Delta[\text{POC}]^\ddagger$	$\Delta - \Delta^{14}\text{C}_{\text{POC}}^\S$	$\Delta[\text{POC}]^\parallel$		
North Atlantic	-167‰	$3\text{ }\mu\text{g l}^{-1}\text{ }^\ddagger$	$+93\text{‰}$	$28\text{ }\mu\text{g l}^{-1}$	0.84	0.16
North Pacific	-83‰	$2\text{ }\mu\text{g l}^{-1}\text{ }^\P$	$+94\text{‰}$	$31\text{ }\mu\text{g l}^{-1}$	0.95	0.05

Estimates of the relative contributions of margin and surface ocean-derived DOC and suspended POC to the deep central North Atlantic and Pacific gyres. Shown are DOC and suspended POC $\Delta^{14}\text{C}$ and concentration gradients between the deep open North Atlantic and North Pacific Oceans, their respective margins (WNA and ENP, respectively), and modern surface ocean organic carbon. Also shown are the fractions of margin and surface ocean DOC and POC required to maintain the observed steady-state $\Delta^{14}\text{C}$ profiles of DOC and suspended POC in the deep North Atlantic and Pacific central gyres.

* Relative contributions of margin and surface-derived DOC and POC required to maintain the observed average $\Delta^{14}\text{C}$ values of DOC and POC in the deep central North Atlantic and Pacific gyres. The contribution of each component is described by the mass balance equation (here shown for DOC): $x[(\Delta - \Delta^{14}\text{C}_{\text{DOC}})_{\text{margin}}] + y[(\Delta - \Delta^{14}\text{C}_{\text{DOC}})_{\text{surface}}] = (\Delta - \Delta^{14}\text{C}_{\text{DOC}})_{\text{deep central gyre}}$, where x is the margin component, y is the surface component, and $x + y = 1$. Note that by definition $(\Delta - \Delta^{14}\text{C}_{\text{DOC}})_{\text{deep central gyre}} = 0$. For calculating the margin and surface-derived suspended POC components, POC values are substituted for DOC values in the above equation. Average values used for $[\text{DOC}]_{\text{deep central gyre}}$ were 43 and $35\text{ }\mu\text{M}$ for the SS and CNP, respectively; the average value used for $[\text{POC}]_{\text{deep central gyre}}$ was $0.72\text{ }\mu\text{g l}^{-1}$ for both the SS and CNP⁹.

† Mean observed gradients in $\Delta^{14}\text{C}$ values of DOC and POC between margins and deep central gyres.

‡ Mean observed gradients in concentrations of DOC and POC between margins and deep central gyres.

§ Mean observed gradients in $\Delta^{14}\text{C}$ values of DOC and POC derived from recent photosynthetically fixed surface ocean organic carbon and that present in deep central gyres. The average $\Delta^{14}\text{C}$ values of surface-derived, photosynthetically fixed carbon measured at the time each central gyre site was sampled were $+120\text{‰}$ and $+140\text{‰}$ for the Sargasso Sea and central North Pacific, respectively⁹.

|| Mean observed gradients in concentrations of DOC and POC between surface (average of 0–20 m depth for DOC; average of particle maximum zone for POC_{susp} , typically 20–85 m depth) and deep (>1,000 m depth) central gyre waters⁹.

¶ Excludes 0 to $\sim 700\text{ m}$ depth interval.

Excludes shallowest depths sampled where concentrations were up to 30 times greater in margins.

Shallow continental shelf and slope waters may also act as low-salinity conduits of younger terrestrial organic matter (J.E.B., unpublished data, and ref. 18), where margins are affected significantly by rivers and estuaries. However, most of this material must also be degraded in nearshore waters or sequestered in sediments as it does not appear to comprise a significant component of open ocean DOC²⁹ and POC_{susp} seaward of the shelf-slope front. The isotope signatures of DOC and POC_{susp} at the coastal–open ocean boundaries (that is, slope and rise waters) here indicate that this carbon has mainly a non-recent marine origin and is older than organic carbon from the North Atlantic and Pacific central gyres. If this material propagates seaward, possibly along isopycnal surfaces, it may represent a source of old DOC and POC to intermediate and deep waters of the interior ocean.⁴ □

Received 11 September 1997; accepted 21 January 1998.

- Walsh, J., Rowe, G., Iverson, R. & McRoy, C. Biological export of shelf carbon is a sink of the global CO₂ cycle. *Nature* **292**, 196–201 (1981).
- Smith, S. & MacKenzie, F. The ocean as a net heterotrophic system: implications for the carbon biogeochemical cycle. *Global Biogeochem. Cycles* **1**, 187–198 (1987).
- Smith, S. & Hollibaugh, T. Coastal metabolism and the oceanic organic carbon balance. *Rev. Geophys.* **31**, 75–89 (1993).
- Wollast, R. in *Ocean Margin Processes in Global Change* (eds Mantoura, R., Martin, J.-M. & Wollast, R.) 365–381 (Wiley, New York, 1991).
- Biscaye, P., Flagg, C. & Falkowski, P. The Shelf Edge Exchange Processes experiment, SEEP-II: an introduction to hypotheses, results and conclusions. *Deep-sea Res.* **41**, 231–252 (1994).
- Walsh, J., Biscaye, P. & Csanady, G. The 1983–1984 Shelf Edge Exchange Processes (SEEP)-I experiment: hypotheses and highlights. *Cont. Shelf Res.* **8**, 435–456 (1988).
- Williams, P. & Druffel, E. Radiocarbon in dissolved organic matter in the central North Pacific Ocean. *Nature* **330**, 246–248 (1987).
- Bauer, J., Williams, P. & Druffel, E. ¹⁴C activity of dissolved organic carbon fractions in the north-central Pacific and Sargasso Sea. *Nature* **357**, 667–670 (1992).
- Druffel, E., Williams, P., Bauer, J. & Ertel, J. Cycling of dissolved and particulate matter in the open ocean. *J. Geophys. Res.* **97**, 15,639–15,659 (1992).
- Trumbore, S. & Druffel, E. in *Role of Nonliving Organic Matter in the Earth's Carbon Cycle* (Zepp, R. G. & Sonntag, C.) 7–22 (Wiley and Sons, New York, 1995).
- Druffel, E., Bauer, J., Williams, P., Griffin, S. & Wolgast, D. Seasonal variability of radiocarbon in particulate organic carbon in the northeast Pacific. *J. Geophys. Res.* **101**, 20,543–20,552 (1996).
- Bauer, J. E., Druffel, E., Williams, P. M., Wolgast, D. & Griffin, S. Temporal variability in dissolved organic carbon and radiocarbon in the eastern north Pacific Ocean. *J. Geophys. Res.* (in the press).
- Williams, P., Robertson, K., Souter, A., Griffin, S. & Druffel, E. Isotopic signatures (¹³C, ¹⁵N) as tracers of sources and cycling of soluble and particulate organic matter in the Santa Monica Basin. *Prog. Oceanogr.* **30**, 253–290 (1992).
- Smith, K., Kaufman, R. & Baldwin, R. Coupling of near-bottom pelagic and benthic processes at abyssal depths. *Limnol. Oceanogr.* **39**, 1101–1118 (1994).
- Stuiver, M., Pearson, G. W. & Braziunas, T. Radiocarbon age calibration of marine samples back to 9000 cal yr B.P. *Radiocarbon* **28**, 980–1021 (1986).
- Anderson, R., Rowe, G., Kemp, P., Trumbore, S. & Biscaye, P. Carbon budget for the mid-slope depocenter of the Middle Atlantic Bight. *Deep-sea Res.* **41**, 669–703 (1994).
- Churchill, J., Wirick, C., Flagg, C. & Pietrafesa, L. Sediment resuspension over the continental shelf east of the Delmarva Peninsula. *Deep-sea Res.* **41**, 341–363 (1994).
- Guo, L., Santschi, P., Cifuentes, L. & Trumbore, S. Cycling of high-molecular-weight dissolved organic matter in the Middle Atlantic Bight as revealed by carbon isotopic (¹³C and ¹⁴C) signatures. *Limnol. Oceanogr.* **41**, 1242–1252 (1996).
- Tanaka, N., Monaghan, M. & Turekian, K. ¹⁴C balance for the Gulf of Maine, Long Island Sound and the northern Middle Atlantic Bight: evidence for the extent of the Antarctic Intermediate Water contribution. *J. Mar. Res.* **48**, 75–87 (1990).
- Bauer, J., Spies, R., Vogel, J., Nelson, D. & Southon, J. Natural ¹⁴C evidence of fossil carbon cycling in sediments of a marine hydrocarbon seep off California. *Nature* **348**, 230–232 (1990).
- Carlson, C. & Ducklow, H. Dissolved organic carbon in the upper ocean of the central Equatorial Pacific, 1992: daily fine-scale vertical variations. *Deep-sea Res.* **42**, 639–656 (1995).
- Cherrier, J., Bauer, J. & Druffel, E. Utilization and turnover of labile dissolved organic matter by bacterial heterotrophs in eastern North Pacific waters. *Mar. Ecol. Prog. Ser.* **139**, 267–279 (1996).
- Hedges, J. et al. Organic ¹⁴C in the Amazon River system. *Science* **231**, 1129–1131 (1986).
- Bauer, J., Reimers, C., Druffel, E. & Williams, P. Isotopic constraints on carbon exchanges between deep ocean sediments and seawater. *Nature* **373**, 686–689 (1995).
- Jahnke, R., Reimers, C. & Craven, D. Intensification of recycling of organic matter at the sea floor near ocean margins. *Nature* **348**, 50–54 (1990).
- Knauss, J. A. *Introduction to Physical Oceanography* (Prentice-Hall, Englewood Cliffs, NJ, 1978).
- Carlson, C. A., Ducklow, H. W. & Michaels, A. F. Annual flux of dissolved organic carbon from the euphotic zone in the northwestern Sargasso Sea. *Nature* **371**, 405–408 (1994).
- Kirchman, D., Suzuki, Y., Garside, C. & Ducklow, H. High turnover rates of dissolved organic carbon during a spring phytoplankton bloom. *Nature* **352**, 612–614 (1991).
- Opsahl, S. & Benner, R. Distribution and cycling of terrigenous dissolved organic matter in the ocean. *Nature* **386**, 480–482 (1997).
- Vogel, J., Nelson, E. & Southon, J. ¹⁴C background levels in an accelerator mass spectrometry system. *Radiocarbon* **29**, 323–333 (1987).
- Stuiver, M. & Polach, H. Discussion: reporting of ¹⁴C data. *Radiocarbon* **19**, 355–363 (1977).

Acknowledgements. We thank D. Wolgast and S. Griffin for their expert assistance in all phases of this work; K. Smith for ship time support of our studies in the eastern N. Pacific; M. Kashgarian, J. Southon and I. Proctor of the Center for AMS at Lawrence Livermore National Laboratory for ¹⁴C analyses; and J. Hedges and R. Jahnke for reviewing earlier versions of the manuscript. This work was supported by the Ocean Margins Program of the US Department of Energy and the Chemical Oceanography Program of the US NSF.

Correspondence and requests for materials should be addressed to J.E.B. (e-mail: bauer@vims.edu).

The deep structure of a sea-floor hydrothermal deposit

Robert A. Zierenberg¹, Yves Fouquet², D. J. Miller³, J. M. Bahr⁴, P. A. Baker⁵, T. Bjerkgård⁶, C. A. Brunner⁷, R. C. Duckworth⁸, R. Gable⁹, J. Gieskes¹⁰, W. D. Goodfellow¹¹, H. M. Gröschel-Becker¹², G. Guérin¹³, J. Ishibashi¹⁴, G. Iturrino¹⁵, R. H. James¹⁶, K. S. Lackschewitz¹⁷, L. L. Marquez¹⁸, P. Nehlig¹⁹, J. M. Peter²⁰, C. A. Rigsby²¹, P. Schultheiss²², W. C. Shanks III²³, B. R. T. Simoneit²⁴, M. Summitt²⁵, D. A. H. Teagle²⁶, M. Urbat²⁷ & G. G. Zuffa²⁸

Hydrothermal circulation at the crests of mid-ocean ridges plays an important role in transferring heat from the interior of the Earth^{1–3}. A consequence of this hydrothermal circulation is the formation of metallic ore bodies known as volcanic-associated massive sulphide deposits. Such deposits, preserved on land, were important sources of copper for ancient civilizations and continue to provide a significant source of base metals (for example, copper and zinc)^{4–6}. Here we present results from Ocean Drilling Program Leg 169, which drilled through a massive sulphide deposit on the northern Juan de Fuca spreading centre and penetrated the hydrothermal feeder zone through which the metal-rich fluids reached the sea floor. We found that the style of feeder-zone mineralization changes with depth in response to changes in the pore pressure of the hydrothermal fluids and discovered a stratified zone of high-grade copper-rich replacement mineralization below the massive sulphide deposit. This copper-rich zone represents a type of mineralization not previously observed below sea-floor deposits, and may provide new targets for land-based mineral exploration.

Transects of boreholes drilled on the Ocean Drilling Program (ODP) Leg 169 defined the extent and composition of the Bent Hill massive sulphide (BHMS) deposit in Middle Valley on the northern Juan de Fuca spreading centre (Fig. 1), and is the first geometric characterization of a complete sea-floor hydrothermal system. A 500-m-deep hole through the apex of the massive sulphide deposit provided a complete vertical section of the hydrothermal system, penetrating the feeder zone, the underlying sediment column and the uppermost volcanic basement.

Middle Valley is the northern extension of the Endeavor segment of the Juan de Fuca ridge. Due to its proximity to the continent and

¹ Department of Geology, University of California-Davis, Davis, California 95616, USA. ² Institut Français de Recherche pour l'Exploitation de la Mer, Centre de Brest, DRO/GM, BP 70 29980 Plouzané Cedex, France. ³ Ocean Drilling Program, Texas A&M Research Park, 1000 Discovery Drive, College Station, Texas 77845, USA. ⁴ Department of Geology and Geophysics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. ⁵ Department of Geology, Duke University, Durham, North Carolina 27708, USA. ⁶ Norges Geologiske Undersøkelse, 7002 Trondheim, Norway. ⁷ Institute of Marine Sciences, University of Southern Mississippi, Stennis Space Center, Mississippi 39529, USA. ⁸ Department of Earth Sciences, James Cook University, Townsville, Queensland 4811, Australia. ⁹ Département Hydrologie et Transefers, Bureau de Recherches Géologiques et Minières, 45060 Orleans Cedex 2, France. ¹⁰ Scripps Institution of Oceanography, University of California-San Diego, La Jolla, California 92093, USA. ¹¹ Geological Survey of Canada, Ottawa, Ontario K1A 0E8, Canada. ¹² Marine Geology and Geophysics, Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, Florida 33149, USA. ¹³ Borehole Research Group, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA. ¹⁴ Laboratory for Earthquake Chemistry, University of Tokyo, Tokyo 113, Japan. ¹⁵ Borehole Research Group, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA. ¹⁶ Department of Geology, University of Bristol, Bristol BS8 1RJ, UK. ¹⁷ Geologisch-Paläontologisches Institut, Christian-Albrechts-Universität Kiel, D24118 Kiel, Germany. ¹⁸ Department of Geological Sciences, Northwestern University, Evanston, Illinois 60208, USA. ¹⁹ Bureau de Recherches Géologiques et Minières, 45060 Orleans Cedex 2, France. ²⁰ Geological Survey of Canada, Ottawa, Ontario K1A 0E8, Canada. ²¹ Department of Geology, East Carolina University, Greenville, North Carolina 27858, USA. ²² GEOTEK Ltd., Northants NN11 5RD, UK. ²³ US Geological Survey, Denver Federal Center MS 973, Denver, Colorado 80225, USA. ²⁴ College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA. ²⁵ School of Oceanography, University of Washington, Seattle, Washington 48109, USA. ²⁶ Geological Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA. ²⁷ Geologisches Institut, Universität zu Köln, D-50674 Köln, Germany. ²⁸ Dipartimento di Scienze della Terra, Università di Bologna, 40127 Bologna, Italy.

increased sediment supply during low stands of sea level caused by Pleistocene glaciation, the spreading centre is buried by continentally derived turbiditic sediments. Active extension and volcanism were contemporaneous with sediment infilling of the rift, resulting in formation of a basaltic-sill/sediment complex that forms the uppermost oceanic crust, in contrast to the extensive basaltic flows at sediment-free spreading centres^{7,8}. High permeability in the sill-sediment complex allows vigorous hydrothermal convection and has resulted in the formation of a relatively isothermal hydrothermal reservoir capped by several hundred metres of sediment with low cross-stratal permeability⁹. Active hydrothermal discharge (Dead Dog vent field, 'Ore Drilling Program' (ODP) mound) and massive sulphide deposits that formed at sites of recent (<300 kyr ago) high-temperature discharge (BHMS, ODP mound) occur within Middle Valley^{7,10}.

The BHMS deposit is located on the south flank of a sediment hill (Bent Hill, Fig. 1) approximately 400 m in diameter and 50 m high that was elevated above the surrounding turbidite plane by the intrusion of late Pleistocene to Holocene basalts⁷. The BHMS deposit is a 35-m-high mound topped by oxidized sulphide rubble and lapped by turbiditic sediment. A slightly younger massive sulphide deposit (ODP mound) is exposed ~300 m to the south (Fig. 1), where before drilling, a single vent was discharging 264°C hydrothermal fluid¹¹.

Coring of the BHMS deposit during Leg 139 (holes 856C-H) indicated that the bulk of the sulphide was deposited above the sea floor, with little evidence for sulphide veining or replacement of

sediment⁷. The primary hydrothermal precipitate was pyrrhotite with less abundant high-temperature Cu-Fe sulphide (isocubanite) and Zn sulphide (sphalerite ± wurtzite)^{7,12,13}. Much of the deposit has been recrystallized to pyrite ± magnetite, and base-metal sulphide showed textural evidence for dissolution, late-stage veining and replacement.

Hole 856H was re-entered on Leg 169 and drilled to a depth of 500 m below sea floor (m.b.s.f.). This core provides a complete vertical section through the massive sulphide deposit and underlying feeder zone, and continues through the sediment into the uppermost oceanic crust (Fig. 2).

Based on both the recovered core and downhole geophysical logging, the massive sulphide deposit has a minimum thickness of 100 m. The massive sulphide zone is a mound with relatively steep flanks and a subhorizontal base (Fig. 2). The base of the massive sulphide lens is interpreted to be a time horizon marking the onset of vigorous hydrothermal venting. A conservative estimate of the size of the massive sulphide mound is approximately 8.8×10^6 tons, but this estimate does not include the significant Cu-rich feeder-zone mineralization deposited below the sea floor or the mineralization associated with ODP mound, which may be of approximately equal size.

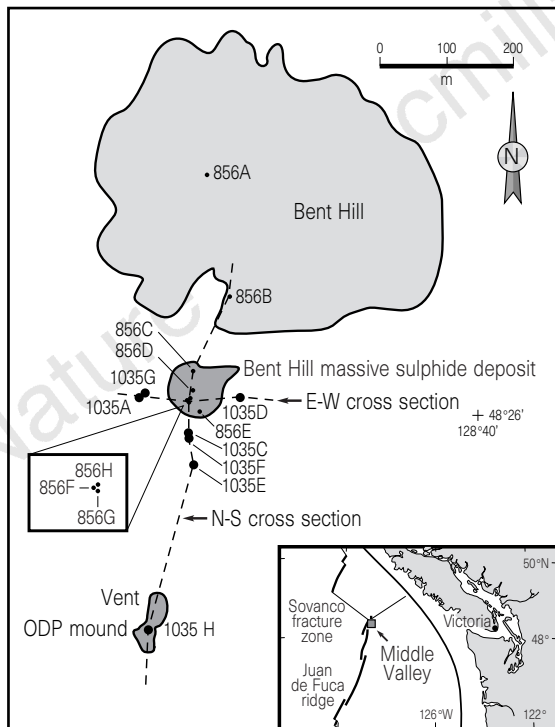


Figure 1 Location of Middle Valley (inset) and details of the study area (main figure). Two massive sulphide deposits are exposed at the sea floor to the south of Bent Hill, which is a recently uplifted sediment hill underlain by basaltic intrusions. The boundaries of Bent Hill, the Bent Hill massive sulphide deposit, and the ODP mound are interpreted from SeaMARC 1 side-scan sonar and constrained by piston coring and observations from the submersible *Alvin*.

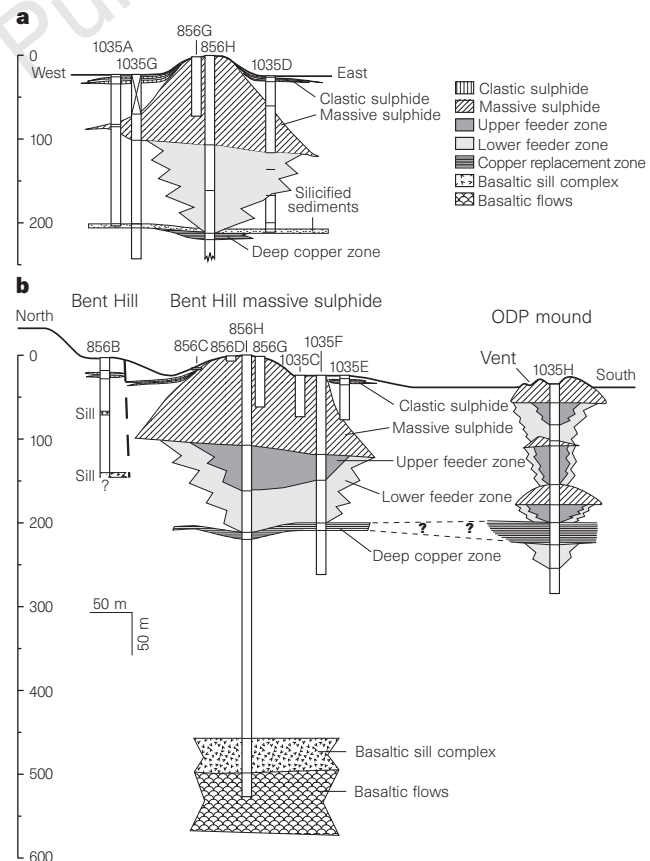


Figure 2 Interpreted cross-sectional views of the distribution of mineralized zones. This information was obtained from SeaMARC1 side-scan sonar, observations from the submersible *Alvin*, piston coring and drill holes from ODP legs 139 (856) and 169 (1035). **a**, East-west section across the Bent Hill massive sulphide deposit. **b**, North-South cross-section of the same area, showing the flank of the Bent Hill sediment hill, Bent Hill massive sulphide deposit and ODP mound. The interpreted structure under the ODP mound is speculative as it is constrained by a single drill hole. The hydrothermal vent at the north end of ODP mound was the only known source of hydrothermal discharge in this area before drilling. Holes 1035F and 1035H were observed to be vigorously venting hydrothermal fluid after drilling.

The lack of interbedded sediment indicates that BHMS deposit formed in an intense episode of sulphide mound building that was rapid relative to the rate of turbidite sedimentation. In contrast, multiple episodes of hydrothermal discharge are clearly evident in the ODP mound (Hole 1035H). Here, three stacked sequences of massive sulphide, each underlain by sediment-hosted feeder-zone mineralization, occur over a depth interval of 210 m (Fig. 2) indicating multiple episodes of hydrothermal activity. The lower sulphide horizons are locally coarse grained due to recrystallization by hydrothermal fluids that formed overlying sulphide lenses. The material recovered from ODP mound is much higher grade (2.9–51% Zn, 0.12–0.62% Cu) than massive sulphide from BHMS¹⁴.

One of the main accomplishments of Leg 169 was the first successful recovery of feeder-zone mineralization underlying a sea-floor massive sulphide deposit, best represented in cores from 100 to 210 m.b.s.f. in Hole 856H (Fig. 2). This interval has been subdivided into three subunits based on the style and intensity of mineralization and alteration. The upper 45 m is intensely mineralized by subvertical veins of intergrown isocubanite-chalcopyrite and pyrrhotite, with vein density decreasing downcore. The host turbidites are altered to chlorite, quartz and fine-grained rutile and titanite. Veins range from >8 cm to <1 mm, with the thickest veins showing crack-seal textures indicative of multiple dilation due to fluid overpressure followed by mineral precipitation (Fig. 3).

The interval from 145 to 200 m.b.s.f. is less intensely veined. The relative proportion of subhorizontal veins and bedding parallel disseminated sulphide increases progressively downcore, with many subvertical veins branching off into subhorizontal sulphide impregnations of the more permeable horizons.

Core recovered between 200 to 210 m.b.s.f. is again intensely mineralized, containing up to 50 vol.% sulphide minerals, predominantly isocubanite containing coarse exsolution lamellae of iron-rich chalcopyrite. Pyrrhotite is much less abundant than in the overlying veins and other sulphide minerals are present in only trace amounts. A representative sample of high-grade mineralization from this zone contains 16.1 wt % copper¹⁴ and this subunit has been designated the deep copper zone (DCZ). In contrast to the

overlying intervals, sulphide veining is essentially absent. Sulphide mineralization occurs as impregnations and replacement of the host sediments, and is strongly controlled by variation in the original sedimentary textures. Much of the mineralization is developed in medium- to coarse-grained, locally crossbedded, turbiditic sand, and preserves the original sedimentary structures (Fig. 3).

Feeder-zone mineralization is also well developed in Hole 1035F on the south flank of the massive sulphide, but is only weakly developed under the east and west flanks of the deposit. The greater extent of sulphide feeder-zone mineralization north–south is consistent with structural control of fluid flow by rift-parallel faulting. The DCZ in Hole 1035F contains more pyrrhotite and sphalerite than in the centre of the system. High-grade Cu-replacement mineralization (8.0–16.6% Cu) also occurs below ODP mound in approximately the same stratigraphic horizon as encountered below the Bent Hill massive sulphide¹⁴. This style of high-grade, copper-rich replacement mineralization below the feeder-zone was not anticipated before drilling, and similar permeable horizons, hitherto unrecognized below ancient massive sulphide deposits on land, may represent potential new exploration targets.

The transition from dominantly vertical crack-seal veins in the upper part of the feeder zone to subhorizontal mineralization controlled by sedimentary texture at the base of the feeder zone indicates that cyclic overpressure capable of fracturing the rock only occurred near the sea floor. Drilling on the east and west flanks of the deposit in holes 1035D and 1035A encountered a weakly mineralized, highly silicified mudstone horizon at the approximate depth of top of the DCZ horizon that provided an important hydrological control on the high-temperature hydrothermal system that formed the massive sulphide. During periods when the high-permeability pathways represented by the veins were sealed, fluid was forced to flow laterally into the more permeable sandy turbidite units. Conductive cooling of this ponded hydrothermal fluid facilitated silica deposition¹⁵, thus sealing the top of this interval, and the precipitation of isocubanite, which is the least soluble of the sulphide minerals that occur in this deposit.

Holes 1035F and 1035H penetrated through the silicified zone and were observed by a television camera lowered down the drill string to be vigorously venting hydrothermal fluid. The silicified zone must still form an impermeable caprock that prevents fluids from reaching the sea floor, except when the seal is penetrated by fracturing or drilling. The intensity of hydrothermal flow out of the 25-cm diameter borehole was sufficient to expel coarse (>3 cm) drill cuttings of sedimentary rock and finer fragments of massive sulphide (5 mm), more than 10 m above the sea floor. The vigour of this venting suggests that the drilling penetrated into a strongly overpressured zone.

Drilling beneath the feeder zone in Hole 856H (>210 m.b.s.f.) penetrated lithified, but relatively unaltered metasediments to a depth of 432 m.b.s.f., then 40 m of interlayered basaltic sills and metasediment, and finally 30 m of pillow basalts, interpreted to be the top of the igneous basement. Magnetic profiles across the eastern flank of Middle Valley indicate that the transition from the weakly magnetized rocks of the sill-sediment complex to extrusive basalts occurs near Bent Hill¹⁶. It appears that the Bent Hill deposit formed near the transition from normal oceanic crust to sedimented-rift type crust, but only after a period of time sufficient for the accumulation of 350 m of turbidites which formed a thermal blanket and hydrological seal over the more permeable sediment-sill complex and underlying oceanic crust.

A key element in creating massive sulphide deposits as large as those drilled in Middle Valley is the extended focusing of intense hydrothermal discharge. Ridge-parallel normal faulting probably provided high-permeability pathways for channelled discharge at the sea floor in Middle Valley. Basaltic rocks from the ocean crust provided the dominant source of both heat and base metals for these sulphide deposits. However, it is the presence of a relatively

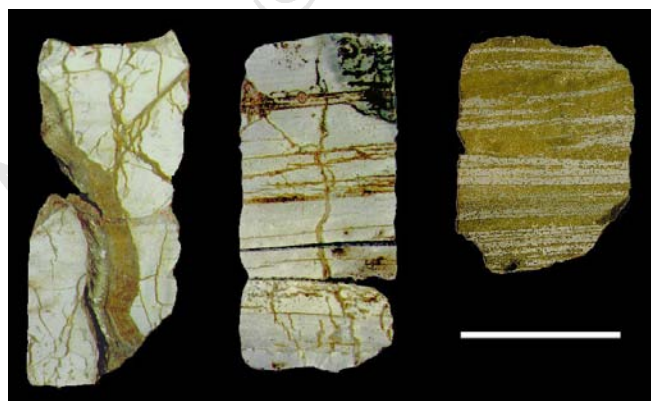


Figure 3 Sections of drill core showing mineralization of the feeder zone and the deep copper zone below the Bent Hill massive sulphide deposit. Left, predominantly vertical crack-seal veins filled with pyrrhotite and isocubanite in altered turbiditic mudstone (856H 24R-1, 50–70 cm, 134 m.b.s.f.). This style of mineralization is characteristic of the upper feeder zone underlying the centre of the hydrothermal upflow zone. Centre, less intense feeder-zone mineralization underlying the south flank of the Bent Hill massive sulphide deposit. Mineralization consists of simple vertical and horizontal veins filled with pyrrhotite, sphalerite and isocubanite in graded fine sand to silt turbidites. Mineralization also occurs as subhorizontal replacement and disseminations along bedding planes (1035F 12R-2 43–55 cm, 112 m.b.s.f.). Right, deep copper zone mineralization in cross-laminated turbiditic sandstone. Replacement of rock by isocubanite mimics original cross-lamination; the matrix is extensively recrystallized to silver-grey coloured chlorite and quartz (856H 31R-1, 99–107 cm, 202 m.b.s.f.).

impermeable sediment blanket in Middle Valley above very young oceanic crust that is the primary control on the style of hydrothermal circulation in this area, resulting in the formation of a sea-floor mineral deposit similar in size and grade to ore deposits mined on land. □

Received 1 August 1997; accepted 29 January 1998.

1. Wolery, T. J. & Sleep, N. H. Hydrothermal circulation and geochemical flux at mid-ocean ridges. *J. Geol.* **84**, 249–275 (1978).
2. Lister, C. R. B. in *The Dynamic Environment of the Ocean Floor* (eds Fanning, K. A. & Manheim, F. T.) 441–470 (Lexington Books, Lexington, 1982).
3. Sleep, N. H. & Wolery, T. J. Egress of hot water from the midocean ridge hydrothermal systems: some thermal constraints. *J. Geophys. Res.* **96**, 2375–2387 (1991).
4. Stanton, R. L. General features of the conformable “pyritic” ore-bodies. *Can. Inst. Mining Metall. Trans.* **63**, 22–27 (1960).
5. Franklin, J. M., Lydon, J. W. & Sangster, D. F. Volcanic-associated massive sulfide deposits. *Econ. Geol., 75th Anniv. Vol.* 485–627 (1981).
6. Hannington, M. D., Jonasson, I. R., Herzog, P. M. & Petersen, S. *Physical and Chemical Processes of Seafloor Mineralization at Mid-ocean Ridges* 115–157 (Geophys. Monogr. 91, Am. Geophys. Union, Washington DC, 1995).
7. Davis, E. E. *et al.* *Proc. ODP Init. Rep.* **139**, 1–1026 (1992).
8. Davis, E. E. & Villinger, H. Tectonic and thermal structure of the Middle Valley sedimented rift, northern Juan de Fuca Ridge. *Proc. ODP Init. Rep.* **139**, 9–41 (1992).
9. Davis, E. E. & Fisher, A. T. On the nature and consequences of hydrothermal circulation in Middle Valley sedimented rift: inferences from geophysical and geochemical observations, Leg 139. *Proc. ODP Sci. Res.* **139**, 695–717 (1994).
10. Goodfellow, W. D. & Franklin, J. M. Geology, mineralogy, and chemistry of sediment-hosted clastic massive sulfides in shallow cores, Middle Valley, northern Juan de Fuca Ridge. *Econ. Geol.* **88**, 2037–2068 (1994).
11. Ames, D. E., Franklin, J. M. & Hannington, M. D. Mineralogy and geochemistry of active and inactive chimneys and massive sulfide, Middle Valley, northern Juan de Fuca Ridge: An evolving hydrothermal system. *Can. Mineral.* **31**, 997–1024 (1993).
12. Krasnov, S., Stepanova, T. & Stepanov, M. Chemical composition and formation of a massive sulfide deposit, Middle Valley, northern Juan de Fuca Ridge (Site 856). *Proc. ODP Sci. Res.* **139**, 353–372 (1994).
13. Duckworth, R. C., Fallick, A. E. & Rickard, E. Mineralogy and sulfur isotopic composition of the Middle Valley massive sulfide deposit, northern Juan de Fuca Ridge. *Proc. ODP Sci. Res.* **139**, 373–385 (1994).
14. Fouquet, Y. *et al.* Middle Valley; Bent Hill area (Site 1035). *Proc. ODP Init. Rep.* **169** (in the press).
15. Janeky, D. J. & Seyfried, W. E. Jr Formation of massive sulfide deposits on oceanic ridge crests: incremental reaction models for mixing between hydrothermal solutions and seawater. *Geochim. Cosmochim. Acta* **48**, 2723–2738 (1984).
16. Currie, R. G. & Davis, E. E. Low crustal magnetization of the Middle Valley sedimented rift inferred from sea-surface magnetic anomalies. *Proc. ODP Sci. Res.* **139**, 19–28 (1994).

Acknowledgements. We thank the staff of the Geological Survey of Canada, particularly J. M. Franklin and E. E. Davis, for sharing data and expertise on the Middle Valley area. We also thank the ODP Engineering staff, and the drilling, ship and scientific staff on-board the D/V *JOIDES Resolution* for their many contributions.

Correspondence and requests for materials should be addressed to R.A.Z. (e-mail: zierenberg@geology.ucdavis.edu).

Megaliths and Neolithic astronomy in southern Egypt

J. McKim Malville, Fred Wendorf*, Ali A Mazar† & Romauld Schild‡

Department of Astrophysical and Planetary Sciences, University of Colorado, Boulder, Colorado 80309, USA

* Department of Anthropology, Southern Methodist University, Dallas, Texas 75275, USA

† Egyptian Geological Survey, Cairo, Egypt

‡ Institute of Archaeology and Ethnology, Polish Academy of Sciences, 00-140 Warsaw

The Sahara west of the Nile in southern Egypt was hyperarid and unoccupied during most of the Late Pleistocene epoch. About 11,000 years ago¹ the summer monsoons of central Africa moved into Egypt, and temporary lakes or playas were formed. The Nabta Playa depression, which is one of the largest in southern Egypt, is a kidney-shaped basin of roughly 10 km by 7 km in area^{2–4}. We report the discovery of megalithic alignments and stone circles next to locations of Middle and Late Neolithic communities at Nabta, which suggest the early development of a complex society. The southward shift of the monsoons in the Late Neolithic age rendered the area once again hyperarid and uninhabitable some 4,800 radiocarbon years before the present (years BP). This well-

determined date establishes that the ceremonial complex of Nabta, which has alignments to cardinal and solstitial directions, was a very early megalithic expression of ideology and astronomy. Five megalithic alignments within the playa deposits radiate outwards from megalithic structures, which may have been funerary structures. The organization of the megaliths suggests a symbolic geometry that integrated death, water, and the Sun. An exodus from the Nubian Desert at ~4,800 years BP may have stimulated social differentiation and cultural complexity in pre-dynastic Upper Egypt.

Pastoralists seem to have entered the Nabta region (Fig. 1 inset) during the summer rainy season beginning ~10,000 years BP. Most of the early sites at Nabta consist of small concentrations of artefacts with one or more hearths, evidence of repeated summer occupation by small family groups. In addition to bones of gazelles, hares, jackals, and small mammals, most of the sites also contain bones of cattle, which may have been used for milk, blood, and transport^{5,6}.

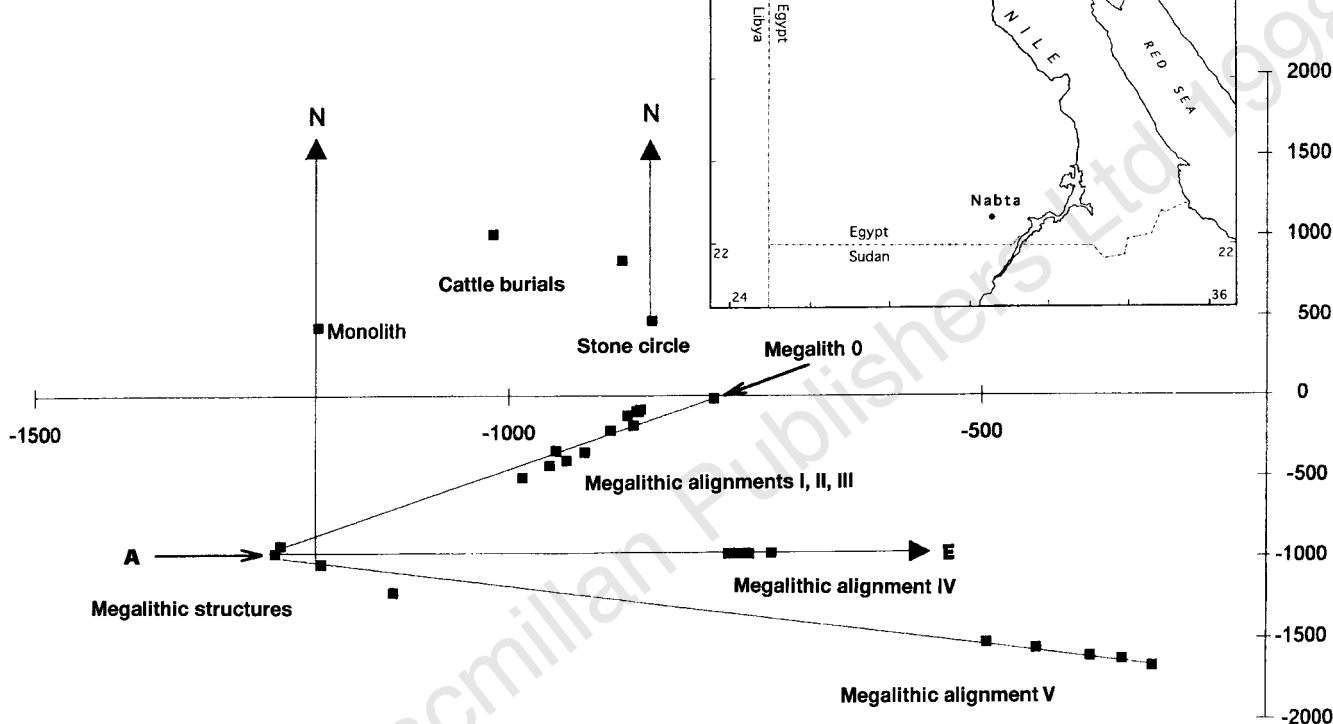
There were three major moist periods in the Holocene epoch in the Eastern Sahara, each of which is documented by massive silt deposits in the seasonal playas, for which we have over 100 radiocarbon dates⁷. These three playa episodes of the Early, Middle, and Late Neolithic ages were separated from each other by periods of hyperaridity, at 7,300–7,100 years BP and 6,700–6,500 years BP, when the water table was lowered to the same or lower levels than those of today. The preceding playa silts were extensively eroded and in some instances sand dunes filled the hollows. The alignments, megalithic structures and sandstone circles were placed in sediments that probably accumulated between 7,000 and 6,700 years BP, at the end of the Middle Neolithic.

These Neolithic settlements reveal repeated occupation over several millennia during the summer rainy season, when there was enough water in the playas for large groups and their animals. At 8,100–8,000 years BP in the Early Neolithic, dates that are well established by a cluster of radiocarbon dates from charcoal and ostrich eggshells, larger communities appeared. One village (E-75-6) contained more than 18 houses, arranged in two (possibly three) straight lines, and deep walk-in wells, which required significant labour investment and control^{8,9}. One well that we excavated was 4 m in width and 3 m deep; the existence of this well may have made it possible for some people to live in the desert throughout the year. The construction of the wells may be the first indication of emerging social control that later made the design and execution of the megalithic complex of the Late Neolithic possible.

Although primarily attracted to the playa for its water and forage, these nomadic groups must have engaged in a variety of activities during summer occupation, such as social bonding, marriage, trade, and ritual. The abundance of cattle remains in the Middle and Late Neolithic settlements is consistent with the ritual traditions of modern pastoralists, who may slaughter cattle to mark socially important events. We excavated two types of cattle tumuli at Nabta. The most common type consists of unshaped blocks of sandstone containing disarticulated bones of one or more cattle. One such tumulus (E-96-1) has yielded a date of 5,500 years BP $\pm$ 160 years, from charcoal in a hearth. The second type of cattle tumulus (E-94-1), which may have marked a place and an event of considerable ideological significance for the group, consisted of an articulated skeleton of a young cow buried in a roofed, clay-lined chamber, which was covered with unshaped sandstone blocks. Wood from the roof of the chamber yielded a radiocarbon date of 6,470 $\pm$ 270 years BP.

Oval clusters of large recumbent slabs constitute the megalithic structures (Fig. 1), which we initially thought might mark high-status burials. However, no firm evidence of human burials was found in any of these features. Although churning clay vertisol would probably have destroyed all buried material except large rocks, the structures may have served primarily as proxy tombs for high-ranking individuals who died on the trail. Excavation and

Figure 1 A plan of the stone structures found in the western portion of the Nabta Playa (scale in metres). A map of Egypt, giving the location of the Nabta Playa, is shown as an inset. True geographic north is indicated. **A** indicates the largest megalithic structure; E is a smaller megalithic structure.



drilling of five of the structures showed that each one was built over a modified table rock, which perhaps functioned symbolically as a cenotaph. We obtained a radiocarbon date of 4,800 years BP $\pm$ 80 from one of the smaller structures (E-96-1; structure E).

Beneath the surface slabs of the largest megalithic structure (E-96-1; structure A) we found a sculptured rock, which has some resemblance to a cow. It was standing upright with its base 2 m below the surface, and its long axis was oriented a few degrees west of north. The rock had been blocked into place by two smaller slabs. Further beneath it, at a depth of 4 m, the shaped table rock had a similar northward orientation.

Excavations of the megaliths contained in the alignments reveal that they are not bedrock material. These slabs, typically measuring 2 m by 3 m, were brought from exposed sandstone, over distances of 0.5 km or more, and then embedded during the Late Neolithic in playa deposits. Megalith 0, with an exposed length of 1.05 m, is shown in Fig. 2 and is the northernmost stone of alignment II. Numerous deflated hearths and Late Neolithic pottery⁹, all of which appear to be contemporaneous with the megalithic alignments, surround the megaliths and cattle tumuli.

The longest series of standing megaliths (megalithic alignments I, II and III; Fig. 1) was originally interpreted² as a single line of megaliths orientated approximately 10° east of north. Our re-evaluation of the alignment indicates that the slabs are organized into three separate lines, which radiate outwards from the largest of the megalithic structures, E-96-1 structure A, with azimuths of 24.3°, 25°, and 28°. During the 1997 season, we combined theodolite and differential global positioning system measurements to map the megaliths, and established the centre of structure A at 22° 30' 29.7" N, 30° 43' 31.2" W. We discovered two additional megalithic alignments, which also radiate out from the vicinity of structure A, with azimuths of 90.02° and 126°. We have not



Figure 2 Megalith 0, the northernmost stone of alignment II (Fig. 1).

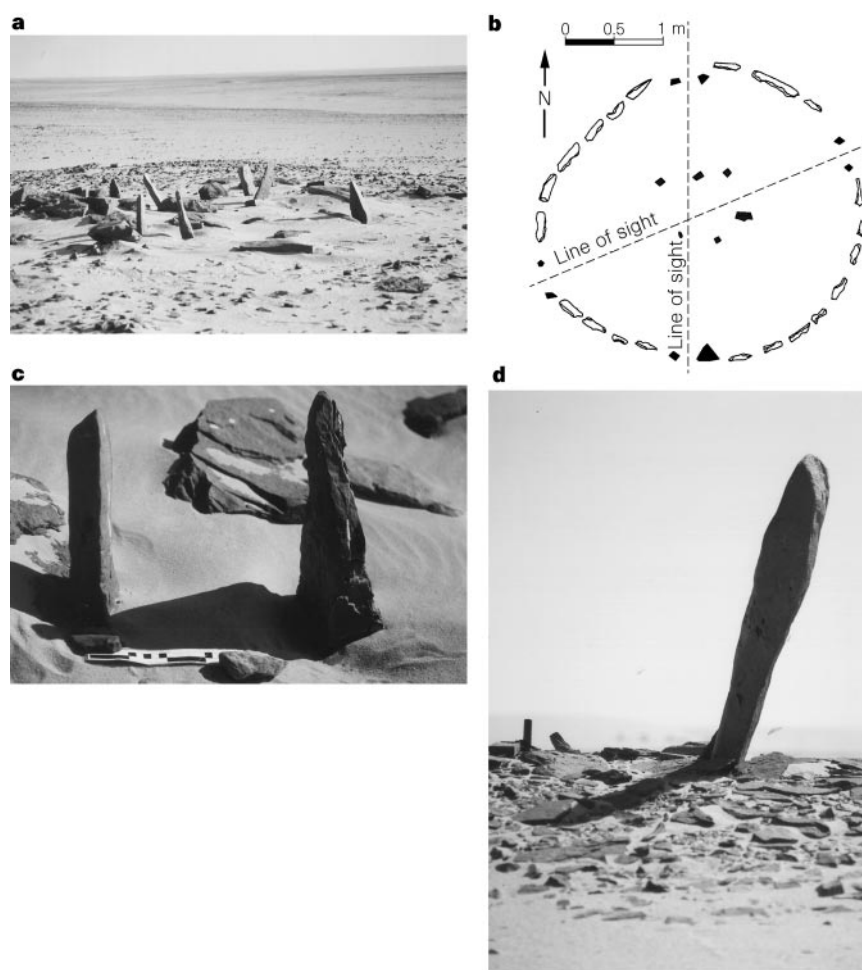


Figure 3 Stone circle and monolith. **a, b**, Stone circle, E-92-9. **b**, The outer eight standing stones that establish sighting slots, in addition to the mix interior standing stones, are shown in black. Recumbent stones are shown in white. **c**, Southwest window of circle in **a, b**. **d**, Standing monolith (height 1 m).

excavated the bases of these megaliths, but they appear to be similarly embedded in playa deposits.

The circle (E-92-9) of small upright and recumbent slabs, with a diameter slightly less than 4 m (Fig. 3a–c), contains four sets of upright slabs, which may have been used for sighting along the horizon. The circle is too small to have functioned as a precise sighting device. The centre lines of the two windows have azimuths of 358° and 62° . Taking into account refraction, we estimate the azimuth of the first gleam of the summer solstice Sun 6,000 years before the present to have been 63.2° , which would have been visible through the slots of the circle. The location on the horizon of the rising Sun close to the summer solstice may have acquired additional significance because of Nabta's proximity to the Tropic of Cancer. At this latitude, the Sun crosses the zenith on two days, approximately three weeks before and after the summer solstice. Vertical structures cast no shadows under the zenith Sun, and within the tropics the day of the zenith Sun is often regarded as a significant event¹⁰.

In addition to the north–south sight-line in the calendar circle, other suggestions of the importance of cardinality are provided by the east–west megalith alignment that extends from structure A and the isolated monolith (Fig. 3d), which lies 1.8° east of north from megalithic structure A. The exposed and buried slabs of structure A, as well as many of the exposed slabs in the other megalithic structures, were also aligned with their long sides approximately north–south.

Although no star was visible at the north celestial pole during most of the occupation of Nabta, north directionality would have been important for nomadic groups navigating across the Sahara. The standing megaliths would have been apt devices to acknowledge the zenith Sun near the onset of the rainy season. Placed in playa

deposits, the megaliths would have been partly submerged in the rising waters of the summer monsoon, and they may have been considered to be ritual markers of the onset of the rainy season. The megalithic complex may have been an expression of interconnections between the Sun, water, death, and the fertile Earth. The unusual standing monolith, either chosen for its shape or intentionally sculptured, is a suggestive symbol of male fertility.

The symbolic richness and spatial awareness seen in the Nabta complex of the Late Neolithic age may have developed from adaptation by nomadic peoples to the stress of survival in the desert. The ceremonial complex could not be more recent than the onset of hyperaridity in the region around 4,800 years BP, suggesting that the astronomy and ceremonialism of Nabta occurred before most of the megalithic features of Europe, Great Britain, and Brittany were established. Within some 500 years after the exodus from Nabta, the step pyramid at Saqqara was constructed, indicating that there was a pre-existing cultural base, which may have originated in the desert of Upper Egypt. An exodus from the Nubian desert at $\sim 5,000$ years BP could have precipitated the development of social differentiation in predynastic cultures through the arrival in the Nile valley of nomadic groups who were better organized and possessed a more complex cosmology. □

Received 14 August 1997; accepted 22 January 1998.

1. Wendorf, F. et al. in *Egypt During the Last Interglacial* (eds Wendorf, F., Close, A. E. & Schild, R.) 552–573 (Plenum, New York, 1993).
2. Wendorf, F., Close, A. E. & Schild, R. Megaliths in the Egyptian Sahara. *Sahara* 5, 7–16 (1992–1993).
3. Wendorf, F. & Schild, R. *Prehistory of the Eastern Sahara* (Academic, New York, 1990).
4. Close, A. E. (ed.) *Prehistory of Arid North Africa* (Southern Methodist Univ. Press, Dallas, 1987).
5. Gautier, A. in *Prehistory of Arid North Africa* (ed. Close, A. E.) 163–187 (Southern Methodist Univ. Press, Dallas, 1987).
6. Close, A. E. & Wendorf, F. in *Transitions to Agriculture in Prehistory* (eds Gebauer, A. B. & Price, T. D.) 63–72 (Prehistory, Madison, 1992).

7. Wendorf, F., Schild, R. & Close, A. *Cattle Keepers of the Eastern Sahara* (Publications in Anthropology, Southern Methodist Univ., Dallas, 1984).
8. Wendorf, F. & Schild, R. Nabta Playa during the Early and Middle Holocene. *ANKH* 4/5, 33–45 (1995–1996).
9. Banks, K. M. in *Prehistory of the Eastern Sahara* (eds Wendorf, F. & Schild, R.) 300–315 (Academic, New York, 1990).
10. Aveni, A. F. Tropical archaeoastronomy. *Science* 243, 161–171 (1981).
11. Burl, A. *From Carnac to Callanish: The Prehistoric Stone Rows and Avenues of Britain, Ireland, and Brittany* (Yale Univ. Press, New Haven, 1993).

Acknowledgements. This paper is based upon research carried out by the Combined Prehistoric Expedition, which is jointly sponsored by the Southern Methodist University, the Institute of Archaeology and Ethnology, the Polish Academy of Sciences, and the Geological Survey of Egypt. Fieldwork was partly supported by the grants from the US National Science Foundation. We thank the Egyptian Antiquities Organization and A. Radwan for support and assistance.

Correspondence and requests for materials should be addressed to J.M.M. (e-mail: malville@spot.colorado.edu).

Inbreeding and extinction in a butterfly metapopulation

Ilik Saccheri*, Mikko Kuussaari*, Maaria Kankare*, Pia Vikman*, Wilhelm Fortelius† & Ilkka Hanski*

* Department of Ecology and Systematics, Division of Population Biology, PO Box 17, 00014 University of Helsinki, Finland

† Tvärminne Zoological Station, University of Helsinki, 10900 Hanko, Finland

It has been proposed that inbreeding contributes to the decline and eventual extinction of small and isolated populations^{1,2}. There is ample evidence of fitness reduction due to inbreeding (inbreeding depression) in captivity^{3–7} and from a few experimental^{8,9} and observational field studies^{10,11}, but no field studies on natural populations have been conducted to test the proposed effect on extinction. It has been argued that in natural populations the impact of inbreeding depression on population survival will be insignificant in comparison to that of demographic and environmental stochasticity^{12,13}. We have now studied the effect of inbreeding on local extinction in a large metapopulation¹⁴ of the Glanville fritillary butterfly (*Melitaea cinxia*)¹⁵. We found that extinction risk increased significantly with decreasing heterozygosity, an indication of inbreeding⁶, even after accounting for the effects of the relevant ecological factors. Larval survival, adult longevity and egg-hatching rate were found to be adversely affected by inbreeding and appear to be the fitness components underlying the relationship between inbreeding and extinction. To our knowledge, this is the first demonstration of an effect of inbreeding on the extinction of natural populations. Our results are particularly relevant to the increasing number of species with small local populations due to habitat loss and fragmentation¹⁶.

The Glanville fritillary metapopulation on the Åland islands in southwest Finland is well suited to the study of factors affecting population extinction^{15,17,18}. This metapopulation consists of numerous small, more-or-less isolated, local populations breeding on dry meadows with one or both of the larval host plants, *Plantago lanceolata* and *Veronica spicata*. The Glanville fritillary has a yearly life cycle in northern Europe. Adult butterflies mate and females lay eggs in June; caterpillars feed in conspicuous family groups of 50–250 larvae, which facilitates large-scale censusing; caterpillars diapause from August until March, continue feeding in the spring and pupate in May. We have located about 1,600 suitable meadows, ranging from 6 m² to 3 ha in size, within an area of 3,500 km². Autumnal surveys have revealed that larvae were present in 524, 401, 384 and 320 meadows in late summer of 1993, 1994, 1995 and 1996, respectively. Local populations can be very small, often consisting of just one sib-group of larvae, the offspring of one pair of butterflies. Consequently, population turnover rate is high, with an average of 200 extinctions and 114 colonizations observed per year. The number of local populations has declined during the study period, probably because of a sequence of unfavourable summers.

Populations were characterized between 1993 and 1995 in terms of size (number of larval groups) and isolation (distances to and the sizes of neighbouring populations¹⁹). Female butterflies were caught in June 1996 from 42 local populations across Åland (Fig. 1), chosen to include relatively large (≥ 5 larval groups), non-isolated populations (from which 5–10 females were sampled per population), as well as small (< 5 larval groups) and isolated populations (from which two females were usually sampled per population).

Individual heterozygosity was determined at seven polymorphic enzyme loci and one polymorphic microsatellite locus (see Methods). The number of heterozygous loci per female was normally distributed, ranging from zero to seven. Heterozygosity differed significantly among the populations ($P = 0.02$). A significant fraction (19%) of variance in heterozygosity among populations was explained by population size in 1993 and by longitude. Heterozygosity was low in populations that had been small in 1993 and in those in eastern Åland. The latter effect apparently reflects large-scale regional changes in abundance in the past^{18,20}.

Accuracy of heterozygosity as a relative measure of inbreeding is largely dependent on the number and degree of polymorphism of markers used to estimate heterozygosity as well as the magnitude of the differences in inbreeding being measured. The variance in inbreeding among populations is expected to be high in this metapopulation, because there is substantial gene flow in many dense regional networks of local populations²¹, but also close inbreeding in many local populations that are extremely small and quite isolated. Thus, differences in average heterozygosity of local populations, even if based on a limited number of polymorphic loci, should reflect real differences in the degree of inbreeding.

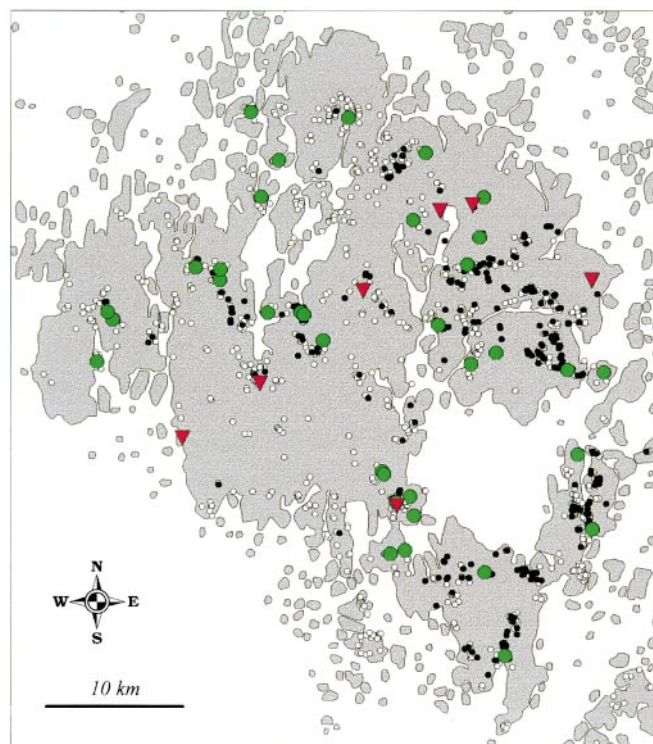


Figure 1 Map of Åland in southwestern Finland showing the locations of the 42 local populations from which adult female butterflies were sampled in summer 1996 (large symbols). All known suitable meadows are shown as small circles, with meadows in which Glanville fritillary larvae were present in autumn 1995 indicated by black circles (and large symbols), and unoccupied meadows by white circles. Of the 42 local populations sampled, the 35 that survived to autumn 1996 (green circles) are distinguished from the seven that went extinct (red triangles).

Seven of the 42 populations went extinct between late summer 1995 and late summer 1996 (Fig. 1). To explain these extinctions, we initially constructed a logistic regression model for the extinction events using population heterozygosity (average number of heterozygous loci per individual) as the explanatory variable (weighted by sample size). The effect of heterozygosity was highly significant ($P > 0.001$). This is unlikely to be the best model for extinction events, however, because many demographic and environmental factors are known to significantly affect extinction risk in the Glanville fritillary^{17,18,22}. Furthermore, heterozygosity may be correlated with these other factors.

We therefore used a model previously developed for extinction events in the entire Åland islands between 1993 and 1994 (ref. 17). In this ('global') model, the risk of extinction increased with decreasing population size ($\log N_{1993}$), with decreasing density of butterflies in the neighbourhood of the focal population (N_{neigh}), with decreasing regional trend in butterfly density (N_{trend}), with decreasing habitat patch size ($\log \text{area}$), and with the incidence of cattle grazing (grazing)¹⁷. We now re-estimated the parameters of this model using the observed extinction events between 1995 and 1996, but excluding the 42 populations from which the genetic data had been collected. The remaining material consisted of 336 local populations in 1995, 185 of which went extinct by 1996. In a logistic regression model, $\log N_{1995}$, N_{trend} and $\log \text{area}$ had a significant

effect on the extinction risk. These are also the three variables that had the strongest effect on extinctions between 1993 and 1994 (ref. 17). Using the parameter values obtained for these three variables, we applied the model to the 42 populations with genetic information, also adding population heterozygosity to the model. The observed extinctions were significantly explained by population heterozygosity in this model ('global model' in Table 1; see also Fig. 2).

An important advantage of this analysis is that the effects of the ecological variables were parameterized using an independent data set. A more accurate prediction can be obtained by directly fitting a model to the actual data from the sample of 42 populations. The 42 populations included more large populations than the remaining 336 populations, which is reflected in the much lower incidence of extinctions in the 42 populations (17% versus 55%). Fitting the original model to the 42 populations, N_{trend} and $\log \text{area}$ had no significant effect, whereas $\log N_{1995}$ and N_{neigh} did. As the abundance of nectar-flowers affects immigration and emigration in the Glanville fritillary²³, we included this variable in the model and found that it significantly influenced extinctions in the predicted direction (increased extinction risk with decreasing abundance of flowers). Finally, we added the estimated population heterozygosity to this model and found that it too had a highly significant effect ('sample model' in Table 1), accounting for 26% of the total deviance. In

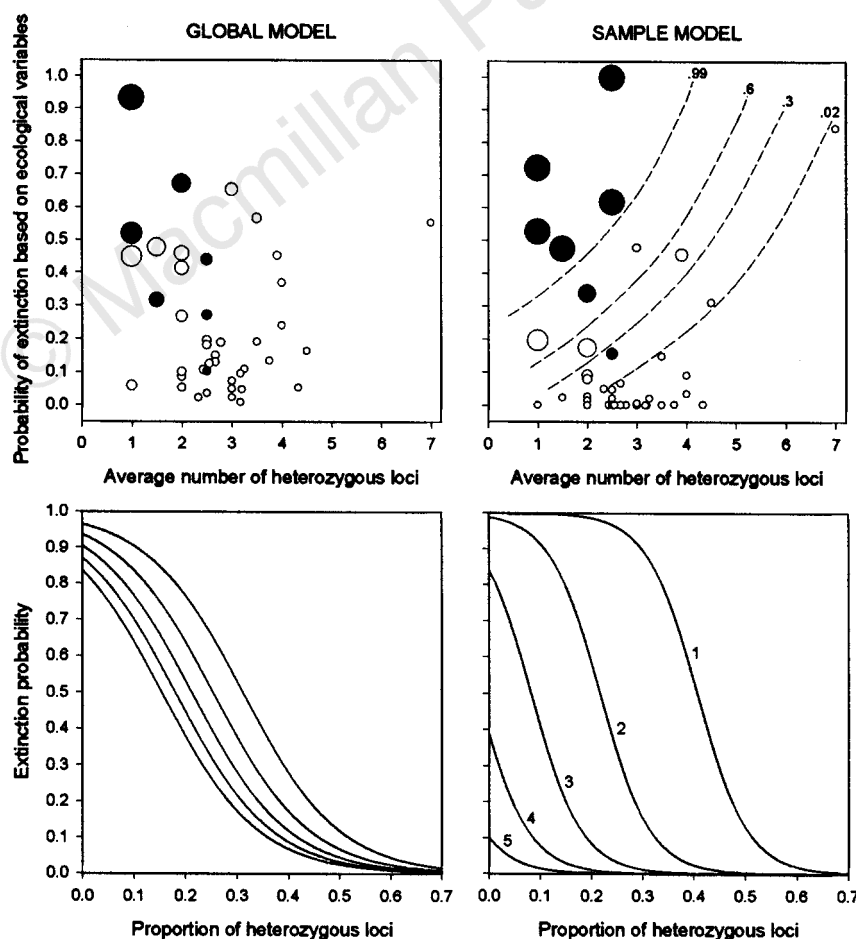


Figure 2 For both global and sample models (Table 1), the upper panels show: (1) the observed average number of heterozygous loci in extinct (black) and surviving (white) populations; (2) the probability of extinction predicted by the models without heterozygosity compared with the observed heterozygosity; (3) the probability of extinction predicted by the full model, including heterozygosity (proportional to circle size). For the sample model, we have drawn appropriate isoclines for the extinction risk predicted by the model, including ecological factors and heterozygosity. These figures illustrate that both the ecological

factors and heterozygosity influence the extinction risk (for statistical analysis, see Table 1). Lower panels show the relationship between the risk of local extinction and heterozygosity predicted by the global and sample models (Table 1). Model predictions are shown for local population sizes of 1–5 larval groups, fixed at the lower quartile value of change in regional density (N_{trend}) and the lower quartile value of meadow area in the global model; and fixed at the lower quartile value of regional density (N_{neigh}) and median flower abundance in the sample model.

Table 1 Two logistic regression models for extinction events in the Glanville fritillary between 1995 and 1996

Variable	Global model				Sample model			
	Coefficient (s.e.)	d.f.	Deviance	P	Coefficient (s.e.)	d.f.	Deviance	P
Constant	5.15 (1.27)				19.89 (9.25)			
Log (N_{1995})	-3.52 (0.61)				-21.90 (10.30)	1	20.02	<0.001
N_{trend}	-3.00 (0.52)							
N_{neigh}					-10.30 (5.62)	1	17.19	<0.001
Log (area)	-0.41 (0.19)							
Flower abundance					-32.30 (17.40)	1	9.68	<0.01
Heterozygosity	-1.33 (0.53)	1	9.61	<0.005	-2.54 (1.12)	1	21.77	<0.001
Residual		40	50.64			37	15.43	
Total		41	60.25			41	84.10	

In the global model the coefficients for variables other than heterozygosity were estimated from the entire metapopulation excluding the 42 populations with genetic data. The global model is the same as previously published¹⁷ for extinctions between 1993 and 1994 but it was re-parameterized for the present material. Variables with a non-significant effect were omitted from the models (the result was essentially the same if all ecological variables used in the previous study¹⁷ were included in the global model). Because the global and sample models contain different explanatory variables, the magnitude of the coefficients for common variables are not directly comparable.

summary, in all three analyses the more heterozygous populations had a lower risk of extinction than the less heterozygous populations.

The relationship between the risk of local extinction and heterozygosity predicted by the global and sample models is shown in Fig. 2. The effect of heterozygosity on extinction risk is most pronounced in small, isolated populations, but even at intermediate levels of isolation (N_{neigh}) extinction risk increases dramatically with decreasing heterozygosity in the smallest populations.

Preliminary findings suggest three major fitness components underlying the observed relationship between heterozygosity and extinction risk in the Glanville fritillary: larval survival, adult longevity and egg hatching. Larval group size shortly after winter diapause was positively associated with maternal heterozygosity (Spearman rank correlation = 0.32, $n = 42$, $P < 0.05$), as was larval weight ($P = 0.004$), suggesting that overall larval viability is enhanced in more heterozygous families (details in Methods). Pupal period was negatively correlated with maternal heterozygosity ($P = 0.04$). In natural populations, a longer pupal period is likely to be associated with reduced survival due to parasitism²⁴.

There was a positive association ($P < 0.05$) between the date females were sampled in the field (from 15 June to 2 July) and their heterozygosity, indicating that females with a shorter life span were more homozygous and inbred. As females are capable of producing a lifetime total of up to seven batches of 50–300 eggs²⁵, which they are constrained to lay on different days when weather conditions are favourable, a small reduction in life expectancy due to inbreeding may have large effects on reproductive output and therefore on population dynamics.

Laboratory studies of the Glanville fritillary from Åland have shown that one generation of brother–sister mating reduces average egg-hatching rate by 24–46% (W.F. *et al.*, unpublished results; see Methods), indicating that the metapopulation has remained sensitive to inbreeding in spite of frequent episodes of local inbreeding in the often very small populations, where sib-mating must occur frequently. In the 42 populations sampled, variance in average egg-hatching rate (determined for one to three egg clutches per female) was significantly larger among populations with low heterozygosity ($\leq$ two heterozygous loci per individual) than in populations with high heterozygosity ($\geq$ three heterozygous loci) (Bartlett's test of equal variances, $P = 0.007$). This was essentially due to a low egg-hatching rate in several populations with low heterozygosity and is an expected outcome of both inbreeding and the interaction of inbreeding and selection^{6,7}.

Detection of the effect of inbreeding on extinction risk was possible in this study by virtue of numerous extinction-prone local populations varying in their degree of inbreeding. Our results suggest that the Glanville fritillary metapopulation maintains a high genetic load, making it susceptible to inbreeding depression. Selection against deleterious recessives exposed by localized inbreeding may be relatively inefficient owing to drift within^{26,27} and gene flow among neighbouring small local populations that carry different

deleterious alleles. This little-studied aspect of metapopulation biology may have far-reaching consequences for the expected persistence times of fragmented populations. The general message from this study is that, although demographic and environmental factors are likely to be the primary determinants of extinction risk, the contribution of inbreeding should not be underestimated, especially in species with a highly fragmented population structure. □

Methods

Measurement of heterozygosity. Cellulose acetate electrophoresis was used to assay polymorphism in the following enzymes: *Ak*, *Got-1*, *Idh-1*, *Pep A*, *Pep D*, *Pgi* and *Pgm*. The microsatellite locus used in this study (CINX22) was cloned from a partial genomic DNA library of *M. cinxia*. The number of alleles and observed heterozygosity for each locus were: 3, 0.29 (*Ak*); 2, 0.18 (*Got-1*); 3, 0.39 (*Idh-1*); 9, 0.57 (*Pep A*); 3, 0.24 (*Pep D*); 7, 0.57 (*Gpi*); 2, 0.3 (*Pgm*); 5, 0.57 (CINX22).

Larval survival and growth. One larval group from each wild-caught female was returned to the meadow of origin to minimize the impact of sampling on population dynamics and to monitor larval survival. A second larval group from most females was placed on a large previously unoccupied meadow on an isolated island to study larval development under common environmental conditions. Larval group size was determined in April 1997, shortly after the caterpillars had come out of winter diapause. The significant correlation between larval group size and heterozygosity refers to 103 groups from 42 populations returned to the meadow from which the mother was caught (correlation was calculated for population averages). Larval group size could not be reliably estimated for groups placed in the common meadow owing to starvation and dispersal in many groups caused by local food shortage. Larval groups that had not been badly affected ($>$ ten larvae survived) in the common meadow were collected and weighed. Linear regression showed a significant positive relationship between the average weight of offspring and maternal heterozygosity: $F_{1,22} = 10.46$, $P = 0.004$, $R^2 = 0.30$. The same larvae were reared to adulthood, providing data on larval growth rate, pupal weight and pupal period (S. Haikola, unpublished results).

Female heterozygosity and sampling date. In the generalized linear model of heterozygosity on sampling date, differences in sampling date among populations were accounted for by fitting population means before adding heterozygosity to the model.

Egg-hatching rate. Egg batches collected from individual females were counted from an enlarged photocopy after they had been spread out on a Petri dish. Following an incubation period of 14–21 days at 22–26 °C and 70–90% relative humidity, a second photocopy was taken from which larvae could be counted. The estimates of the effects of inbreeding on egg-hatching rate reported here come from two laboratory experiments conducted at different times using independent samples of butterflies (W.F. *et al.*, unpublished results). In the first experiment (spring 1997), the average egg-hatching rate in six brother–sister matings, originating from five unrelated families from four distant local populations, was 37%, compared to 69% in 12 replicated pairwise interpopulation crosses. The butterflies used in the second experiment (summer 1997) were the F_2 generation from 13 wild families from three distant populations, which were in some cases the product of brother–sister

matings. Average egg-hatching rate in 64 full-sib matings representing all founder families was 60%, compared to 79% in 48 non-sib crosses within and between populations.

Received 22 September 1997; accepted 4 February 1998.

1. Frankham, R. Conservation genetics. *Annu. Rev. Genet.* **29**, 305–327 (1995).
2. Frankham, R. Inbreeding and extinction: a threshold effect. *Conserv. Biol.* **9**, 792–799 (1995).
3. Charlesworth, D. & Charlesworth, B. Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* **18**, 237–268 (1987).
4. Ralls, K., Ballou, J. D. & Templeton, A. Estimates of lethal equivalents and the cost of inbreeding in mammals. *Conserv. Biol.* **2**, 185–193 (1988).
5. Thornhill, N. W. (ed.) *The Natural History of Inbreeding and Outbreeding: Theoretical and Empirical Perspectives* (University of Chicago Press, Chicago, 1993).
6. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* 4th edn (Longman, Burnt Mill, Harlow, UK, 1996).
7. Saccheri, I. J., Brakefield, P. M. & Nichols, R. A. Severe inbreeding and rapid fitness rebound in the butterfly *Bicyclus anynana* (Satyridae). *Evolution* **50**, 2000–2013 (1996).
8. Jiménez, J. A., Hughes, K. A., Alaks, G., Graham, L. & Lacy, R. C. An experimental study of inbreeding depression in a natural habitat. *Science* **266**, 271–273 (1994).
9. Newman, D. & Pilon, D. Increased probability of extinction due to decreased effective population size: experimental populations of *Clarkia pulchella*. *Evolution* **51**, 354–362 (1997).
10. Keller, L. F., Arcese, P., Smith, J. N. M., Hochachka, W. M. & Stearns, S. C. Selection against inbred song sparrows during a natural population bottleneck. *Nature* **372**, 356–357 (1994).
11. Madsen, T., Stille, B. & Shine, R. Inbreeding depression in an isolated population of adders *Vipera berus*. *Biol. Conserv.* **75**, 113–118 (1996).
12. Caro, T. M. & Laurenson, M. K. Ecological and genetic factors in conservation: a cautionary tale. *Science* **263**, 485–486 (1994).
13. Caughley, G. Directions in conservation biology. *Anim. Ecol.* **63**, 215–244 (1994).
14. Hanski, I. A. & Gilpin, M. E. (eds) *Metapopulation Biology: Ecology, Genetics, and Evolution* (Academic, San Diego, 1997).
15. Hanski, I., Pakkala, T., Kuussaari, M. & Lei, G. Metapopulation persistence of an endangered butterfly in a fragmented landscape. *Oikos* **72**, 21–28 (1995).
16. Lande, R. Genetics and demography in biological conservation. *Science* **241**, 1455–1460 (1988).
17. Hanski, I., Pöry, J., Pakkala, T. & Kuussaari, M. Multiple equilibria in metapopulation dynamics. *Nature* **377**, 618–621 (1995).
18. Hanski, I., Moilanen, A., Pakkala, T. & Kuussaari, M. The quantitative incidence function model and persistence of an endangered butterfly metapopulation. *Conserv. Biol.* **10**, 578–590 (1996).
19. Hanski, I. A practical model of metapopulation dynamics. *Anim. Ecol.* **63**, 151–162 (1994).
20. Hanski, I. Metapopulation dynamics: from concepts and observations to predictive models in *Metapopulation Biology: Ecology, Genetics, and Evolution* (eds Hanski, I. A. & Gilpin, M. E.) 69–91 (Academic, San Diego, 1997).
21. Hanski, I., Kuussaari, M. & Nieminen, M. Metapopulation structure and migration in the butterfly *Melitaea cinxia*. *Ecology* **75**, 747–762 (1994).
22. Lei, G. C. & Hanski, I. Metapopulation structure of *Cotesia melitaeaeurum*, a specialist parasitoid of the butterfly, *Melitaea cinxia*. *Oikos* **78**, 91–100 (1997).
23. Kuussaari, M., Nieminen, M. & Hanski, I. An experimental study of migration in the Glanville fritillary butterfly *Melitaea cinxia*. *J. Anim. Ecol.* **65**, 791–801 (1996).
24. Lei, G. C., Vikberg, V., Nieminen, M. & Kuussaari, M. The parasitoid complex attacking Finnish populations of the Glanville fritillary *Melitaea cinxia* (Lep: Nymphalidae), an endangered butterfly. *J. Nat. Hist.* **31**, 635–648 (1997).
25. Wahlberg, N. *One Day in the Life of a Butterfly; a Study of the Glanville Fritillary Melitaea cinxia* (Thesis, Univ. Helsinki, 1995).
26. Hedrick, P. W. Purging inbreeding depression and the probability of extinction: full-sib mating. *Heredity* **73**, 363–372 (1994).
27. Lynch, M., Conery, J. & Bürger, R. Mutation accumulation and the extinction of small populations. *Am. Nat.* **146**, 489–518 (1995).

Acknowledgements. We thank K. Higgins, L. Keller, R. Lande, M. Lynch, A. Moilanen, R. Nichols, S. van Nouhuys, M. Singer and C. Thomas for their helpful comments on the manuscript.

Correspondence and requests for materials should be addressed to I.S. (e-mail: ilik.saccheri@helsinki.fi).

Cerebellar complex spikes encode both destinations and errors in arm movements

Shigeru Kitazawa^{*,†}, Tatsuya Kimura^{*} & Ping-Bo Yin[§]

^{*} Information Science Division, Electrotechnical Laboratory,

[†] PRESTO and [§] CREST, Japan Science and Technology Corporation, 1-1-4 Umezono, Tsukuba 305-8568, Japan

Purkinje cells of the cerebellum discharge complex spikes, named after the complexity of their waveforms¹, with a frequency of ~1 Hz during arm movements^{1–13}. Despite the low frequency of firing, complex spikes have been proposed to contribute to the initiation of arm movements^{2,7–10} or to the gradual improvement of motor skills^{2,4–6,14–16}. Here we recorded the activity of Purkinje cells from the hemisphere of cerebellar lobules IV–VI while trained monkeys made short-lasting reaching movements (of

~200 milliseconds in duration) to touch a visual target that appeared at a random location on a tangent screen. We examined the relationship between complex-spike discharges and the absolute touch position, and between complex-spike discharges and relative errors in touching the screen. We used information theory to show that the complex spikes occurring at the beginning of the reach movement encode the absolute destination of the reach, and the complex spikes occurring at the end of the short-lasting movements encode the relative errors. Thus, complex spikes convey multiple types of information, consistent with the idea that they contribute both to the generation of movements and to the gradual, long-term improvement of these movements.

Two macaque monkeys were trained to make rapid reaching movements toward a visual target that appeared on a tangent screen, located 200 mm from the eyes, from a button positioned 200 mm below the eyes in the mid-sagittal plane. A trial began when the monkey pressed the button (Start, Fig. 1); a target then appeared (Target, Fig. 1) at a random place in a square target zone (50 × 50 mm, or 80 × 80 mm) on the screen. The monkey had to release the button (Release, Fig. 1) within 240 ms of the appearance of the target and touch the screen within 300 ms of releasing the button (Touch, Fig. 1). The monkey's view of its hand and the target was blocked at the release of the button by liquid-crystal shutters in front of the eyes (Release, Fig. 1). The shutters opened again when the screen was touched (Open, Fig. 1), allowing the monkey to see the target and the final position of its hand for 300 ms. The monkey had to hold the final position of its hand for 900 ms until given a reward (Reward, Fig. 1); the size of the reward was in inverse proportion to the magnitude of the error to encourage accurate reaching.

Figure 2 shows a raster plot (Fig. 2a) and the average discharge frequencies (Fig. 2b) of a Purkinje cell from lobule V. Stable simple spikes and complex spikes were recorded from this Purkinje cell during 1,381 trials. The average discharge frequency of simple spikes, (Fig. 2b) decreased to 40 Hz during the movement, and then sharply increased to 170 Hz at the end of the movement (Touch, time zero). In contrast, complex spikes generally occurred only once during each trial (black dots, Fig. 2a) and the average discharge frequency was less sharply altered during the movement (Fig. 2b). To test whether the sporadic complex-spike discharges of this Purkinje cell ever encode absolute destination of the reaching movement or relative errors, we set three time windows (shown with crossed, diagonal or horizontal lines in Fig. 2b).

During the first time window in the first half of the movement (crosshatched area in Fig. 2b), complex-spike discharges occurred in 133 of the 1,381 trials. Figure 3a shows the absolute position of

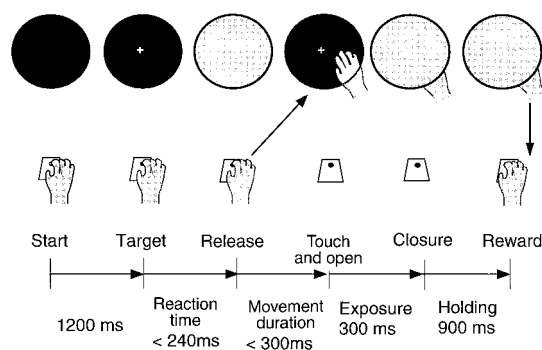


Figure 1 The reaching task. The sequence of events for each trial is shown from left to right. A trial begins when a monkey touches a button. A target beam then appears on the screen. The monkey then releases the button, and touches the target on the screen. At this point, liquid-crystal shutters that were obscuring the monkey's view of its hand are opened for 300 ms, then closed. The monkey holds the position of its hand on the target for 900 ms, and then receives a reward.

touch on the screen in the 1,381 trials (black dots). Results from the 133 trials that showed complex-spike discharges are circled in red. Although the touch positions (black dots) for the 1,381 trials were distributed evenly over the 50×50 mm target zone, the touch positions for the 133 trials that showed complex-spike discharges (red circles) were skewed towards the lower right quadrant of the screen. The number of red circles in each quadrant (counter-clockwise from upper right quadrant: 24, 14, 39, 56) was significantly uneven (χ^2 test, $P < 0.00001$, degrees of freedom d.f. = 3, $\chi^2 = 30.3$). Thus, when complex spikes occurred during the initial

part of the reach, the monkey was four times more likely to touch the fourth quadrant than the second (56 touches out of 133 versus 14 touches). The predictive information associated with the occurrence of complex spikes (I_{cs}) was quantified as the decrease in entropy in the absolute touch position:

$$I_{cs} = \sum_{i=1}^4 - (n_i/N) \log_2(n_i/N) - \sum_{i=1}^4 - (m_i/M) \log_2(m_i/M) \quad (1)$$

where n_i and m_i denote numbers of black dots and red circles in the i th quadrant, and N and M denote total numbers of black dots and red circles, respectively. If the red circles are confined to a particular quadrant, I_{cs} would yield 2 bits, and if they distribute evenly I_{cs} would yield 0 bits. In Fig. 3a, I_{cs} was 0.2 bits. The information was considered 'significant' when the corresponding χ^2 tests yielded a P value < 0.05 , as in the present case ($P < 0.00001$).

Errors of reaching relative to the target in the 1,381 trials (black dots in Fig. 3b, c) were distributed in a gaussian manner around the target, as were the errors in the subset of 133 trials with an early complex spike (results not shown). The P value of the χ^2 test was 0.69, indicating that the complex spike conveyed little information ($I_{cs} = 0.01$) about the reaching errors.

During the second time window, just after touching of the screen (shown with oblique lines in Fig. 2b), complex-spike discharges occurred in 92 of the 1,381 trials. The distribution of touch positions in the 92 trials was almost even over the four quadrants (results not shown; $P = 0.76$, $I_{cs} = 0.01$), but relative reaching errors for the 92 trials were distributed preferentially in the upper left quadrant (red circles in Fig. 3b; $P < 0.0001$, $I_{cs} = 0.19$). During the third time window, 120–220 ms after the touch (shown with horizontal lines in Fig. 2b), 88 complex-spike discharges occurred. Relative errors in the 88 trials were distributed preferentially in the left half (red circles in Fig. 3c; $P < 0.0001$, $I_{cs} = 0.18$), but absolute touch positions were evenly distributed (results not shown; $P = 0.71$, $I_{cs} = 0.01$). Thus, the complex-spike discharges during the second and third time windows conveyed significant information about reaching errors.

These three time windows were not chosen arbitrarily. We calculated the information (I_{cs}) on the touch position and on the relative reaching errors by moving a time window of 100 ms along

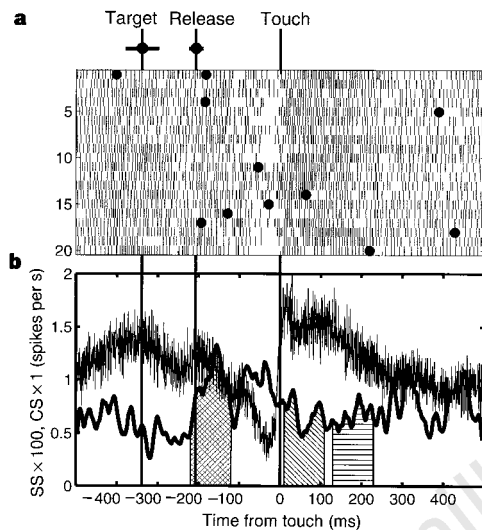


Figure 2 Simple-spike (SS) and complex-spike (CS) activity recorded from a Purkinje cell in lobule V. **a**, A raster plot of SS (vertical bars) and CS (filled dots) activity aligned at the touch (20 trials). **b**, Average discharge frequency (ordinate) of SS activity (thin trace, $\times 100$ Hz) and CS activity (thick trace, $\times 1$ Hz) for 1,381 trials. Bin width is 0.5 ms and the trace for CS activity was smoothened by the moving average method (40-ms window). Mean timing of target appearance and button release is shown by vertical lines, with s.d. (horizontal bars). Three 100-ms time windows are shown with crossed, oblique and horizontal lines, in the lower panel.

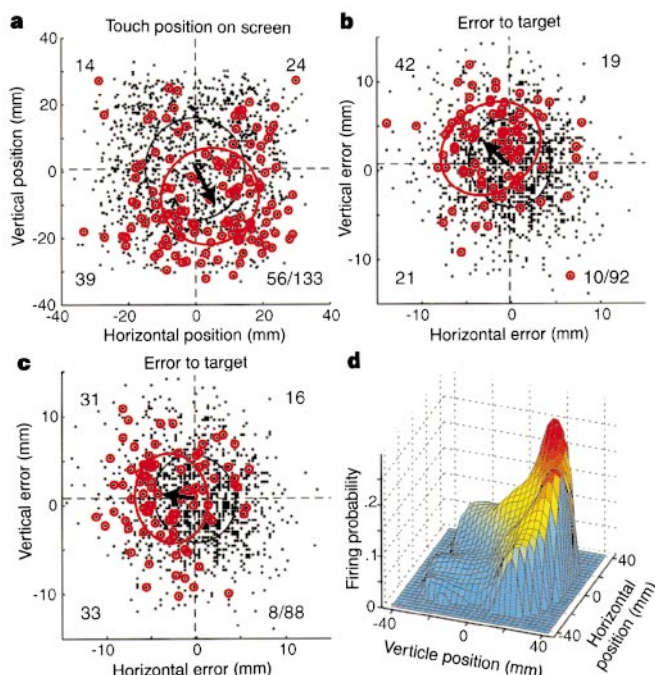


Figure 3 Distribution of touch positions and relative errors. Data are from the same cell used for Fig. 2. **a**, Scatterplots of the touch positions on the screen in all 1,381 trials (black dots), and in the 133 trials (red circles) with complex-spike activity during the first 100-ms window shown in Fig. 2. Horizontal and vertical dashed lines pass through the centre of gravity of the black dots and define the four quadrants. Numbers of red circles in each quadrant are indicated, together with the total number (133) in the fourth quadrant. Black and red ellipses show equidistant points (Mahalanobis distance = 1) from the centre of gravity of the black dots and red circles, respectively. The black arrow connects the centres of gravity of the black dots and the red circles and defines the 'preferred' direction. **b**, **c**, Scatterplots of relative errors to the target (origin) in the 1,381 trials (black dots), and of 92 (**b**) and 88 (**c**) trials (red circles) during which complex-spike activity was seen during the second and the third 100-ms time windows shown in Fig. 2, respectively. **d**, Data shown in **a** transformed into firing probability (z-axis) of complex spikes as a function of touch position. A circular window (radius = 10 mm) was moved with 1-mm steps over the x and y planes in **a**, and the number of red circles within the window was divided by the number of black dots at each position. A median filter (10 $\times$ 10 mm) was also used.

the time axis (abscissa in Fig. 2b) with steps of 10 ms, and calculated the information transmission rate (I_r in bits s^{-1}). I_r can be approximated by the product of I_{cs} and the firing frequency of complex spikes (see Methods). There are three peaks of information (Fig. 4a), corresponding roughly to the three time windows. These three peaks were common in the population of 50 Purkinje cells examined (Fig. 4b). Of these 50 cells, most ($n = 47$) showed changes in simple-spike discharges of $>30\%$ of their maximum discharge rates, when aligned at the onset or at the end of the reaching movements. The other three cells, whose simple spikes were not changed, conveyed no significant information concerning either the touch position or the error. Regarding the information encoding in the complex spikes of these 47 cells, 19 (40%) showed a significant first peak (Fig. 4b), 21 (45%) showed a significant second peak, and 19 (40%) showed a significant third peak. Thirty-four cells (72%) contributed to at least one of the three peaks, and eight (17%) contributed to all of the three peaks (Fig. 4a). In the eight cells that had all three peaks, peak 1 was the largest in four cells (Fig. 4a), and peak 3 was the largest in the other four cells.

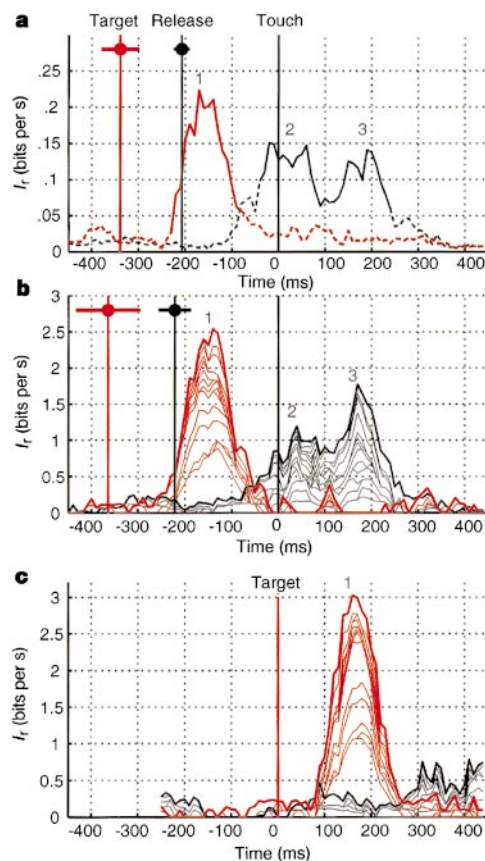


Figure 4 Information transmission rate (I_r) on the touch position (red trace) and the relative error (black trace) encoded by the complex spikes. I_r was calculated using equation (2). **a**, I_r encoded by complex spikes of the same Purkinje cell used for Figs 2, 3. I_r is plotted against the centre of the 100-ms time window that was moved along the time axis (abscissa). Solid and broken traces indicate significant and nonsignificant information, respectively. **b**, **c**, Summation of I_r curves from 50 Purkinje cells aligned on the touch (**b**) and the target appearance (**c**). Nonsignificant information was excluded from the summation. Intervals between the thin lines show the contributions from single cells. This simple summation gives the upper limit of the net information that was encoded by the whole population of cells, as the information from each cell may not be independent. The mean timings for target appearance (red) and release of the button (black) are indicated by vertical lines with horizontal bars (s.d.) in **a**, **b**. Numbers 1–3 in **a–c** indicate the peak numbering used in the text.

For the whole population of cells, the preferred directions shown by cells for each of the three peaks of information were distributed almost randomly (results not shown). However, cells that had more than one peak of information showed a consistent pattern of directional preferences. In the 14 cells that had both peaks for relative errors (peaks 2 and 3), the preferred directions for the two peaks usually fell in the same quadrant (in 12 of the 14 cells; difference: $27 \pm 34^\circ$, mean $\pm$ s.d.) and never fell in opposite quadrants. In contrast, the preferred directions for the errors (peak 2 or 3) fell in the quadrant that was opposite to that for the touch position (peak 1) in 7 of 9 cells (for peak 2) and 6 of 10 cells (peak 3), and none of the two pairs ($n = 19$) fell in the same quadrant (difference: $207 \pm 40^\circ$, mean $\pm$ s.d.).

Thus far, we have only partitioned the work space into quadrants and have considered only the encoding of a partial entity of a vector, the direction of movement. By moving a circular window (radius = 10 mm) in 1-mm steps over x and y planes, such as those in Fig. 3d, we were able to map the probability of complex-spike discharge with a higher spatial resolution and thereby show graded information about the magnitude of movements. This analysis showed that the probability of complex-spike discharge increased progressively up to 0.25 as the position shifted in the preferred direction on the touch screen (Fig. 3d). Similar gradations in the probability of discharge were also seen with the relative errors (data not shown). Thus, the complex-spike discharges encode vectors with both magnitude and direction.

Peak 1 became taller when we aligned the summation of I_r curves with the appearance of the target (Fig. 4c, red line), with an onset 80 ms after the appearance of the target, consistent with visual mediation. The earlier peak of information about the relative errors (peak 2, Fig. 4b) cannot have been caused by visual input because vision was blocked by the shutters (Fig. 1) until the screen was touched (time zero), so this peak presumably resulted from somatosensory feedback and/or efference copy. Because of its latency after the touch (when visual error information was first made available), peak 3 (Fig. 4b) also was probably visual in origin and linked to the error at the touch.

Although analysis indicates that complex spikes during peaks 1 and 3 seem to encode different parameters, it is likely that the brain interprets both types of spike as visual motor error, one at the start and the other at the end of the trial, especially as these two peaks tend to show inverse spatial tuning. Only by blocking any view of the hand during reaching were we able to uncover the early visual encoding of touch position (peak 1) and the early non-visual encoding of errors (peak 2), but, in so doing, we excluded any possibility that the complex spikes could signal visual errors during reaching.

What are the roles of the complex spikes that encoded these three peaks of information at different stages of movement? The first peak of information may encode the absolute destination at the beginning of the movement and so might act as a movement-error signal (in spatial coordinates) helping to guide the hand towards the target^{2,7–13}. The second and third peaks of information may encode the relative error at the end of the reach; this is too late to affect the ongoing ballistic movement, but might help to improve subsequent movements, that is, to refine motor skills^{2,4–6,14–16}. Thus, we suggest that signals for the 'real-time' control and 'learning' of movements¹⁷ are both encoded by complex-spike discharge. □

Methods

Task procedures. We used two monkeys (*Macaca fuscata*, 7.5 kg and 8.5 kg). The monkeys' reaching task is described above and shown in Fig. 1. Eye movements were not controlled during the task. Video pictures showed that the tip of the middle finger touched the screen first in most trials. The animals did not use their palms. The reward was determined by the first position of touch, that is, the position of touch by his middle finger in most trials. The touch screen, liquid-crystal shutters, and reaching task were similar to those used in human experiments¹⁸.

Recording and histological identification of Purkinje-cell locations. After training, a chronic recording chamber was implanted on the skull ipsilateral to the trained arm. Extracellular action potentials of single Purkinje cells were recorded by a glass-coated tungsten microelectrode. Purkinje cells were identified by the presence of spontaneous complex-spike discharge. We recorded as many trials as possible from all Purkinje cells that were encountered, irrespective of their discharge characteristics. The discharges of 161 Purkinje cells were recorded from three hemispheres of the two animals during the reaching task. Simple and complex spikes were discriminated online using a PC-based spike discriminator that applies a principal-component analysis to each spike form¹⁹. Discriminated trains of simple and complex spikes in each trial were stored on a PC with the timings of the events. In each hemisphere, several penetration tracks were marked by lesions made by passing a direct current (10 μ A for 20 s) through the tip of the recording microelectrode. At the end of the recording sessions, the first animal was anaesthetized and perfused with paraformaldehyde. Serial sections of the cerebellum were made. Using the lesions as a reference, recording sites from the first animal ($n = 134$) were all found to be in the intermediate and lateral hemispheres of lobules IV–VI. Data from the second animal are still being recorded but X-ray pictures indicate that the recording locations may be similar to those in the first animal.

Data analysis. Stored data from simple-spike and complex-spike discharges were analysed offline on a workstation (Sun, Ultra 2) using MATLAB. We judged that simple and complex spikes were recorded from isolated Purkinje cells when we confirmed a complete pause in simple spikes for 10 ms after complex-spike onset. Of 161 Purkinje cells, 50 yielded enough data for further analysis in that their simple-spike and complex-spike discharges were recorded for more than 300 trials (33,041 trials in total). Information analysis was applied to each of the 50 cells.

In the analysis, the predictive information available with an occurrence of complex spikes (I_{cs}) was estimated from the decrease in entropy as in equation (1). The amount of information carried by the 'lack' of complex spikes ($I_{\bar{cs}}$) was defined as $I_{\bar{cs}} = \sum_{i=1}^N - (n_i/N) \log_2(n_i/N) - \sum_{i=1}^N - ((n_i - m_i)/(N - M)) \times \log_2((n_i - m_i)/(N - M))$, where the notation is the same as in equation (1). The information carried by the complex spikes (I) was defined as:

$$I = p \cdot I_{cs} + (1 - p) \cdot I_{\bar{cs}} \quad (2)$$

where p is the probability of complex-spike occurrence during the time window under consideration. We further defined the information transmission rate, I_r , as $I_r = I/w$, where w denotes the width of the time window. In equation (2), $I_{\bar{cs}}$ was negligibly small, and p is approximated by the product of firing frequency (f) and w as $p \approx fw$. Hence $I_r \approx f \cdot I_{cs}$.

We used time windows of 100, 150 and 200 ms for those cells that were recorded for more than 600 ($n = 26$), 450 ($n = 13$) and 300 ($n = 11$) trials, respectively. The longer time windows were adopted for cells recorded for fewer trials to allow use of the χ^2 test.

The axis that defined the four quadrants was rotated with a step of 10° , and the calculation of the information and P values was repeated nine times. The average of the nine calculations was used for the calculation of I_r , and for the judgement of significance in the χ^2 test.

Received 17 October 1997; accepted 7 January 1998.

1. Thach, W. T. Discharge of Purkinje and cerebellar nuclear neurons during rapidly alternating arm movements in the monkey. *J. Neurophysiol.* **31**, 785–797 (1968).
2. Simpson, J. I., Wylie, D. R. & De Zeeuw, C. I. On climbing fiber signals and their consequence(s). *Behav. Brain Sci.* **19**, 384–398 (1996).
3. Thach, W. T. Discharge of cerebellar neurons related to two maintained postures and two prompt movements. II. Purkinje cell output and input. *J. Neurophysiol.* **33**, 537–547 (1970).
4. Gilbert, P. F. C. & Thach, W. T. Purkinje cell activity during motor learning. *Brain Res.* **128**, 309–328 (1977).
5. Keating, J. G. & Thach, W. T. Nonclock behavior of inferior olive neurons: interspike interval of Purkinje cell complex spike discharge in the awake behaving monkey is random. *J. Neurophysiol.* **73**, 1329–1340 (1995).
6. Thach, W. T. On the specific role of the cerebellum in motor learning and cognition: clues from PET activation and lesion studies in man. *Behav. Brain Sci.* **19**, 411–431 (1996).
7. Mano, N., Kanazawa, I. & Yamamoto, K. Complex-spike activity of cerebellar P-cells related to wrist tracking movement in monkey. *J. Neurophysiol.* **56**, 137–158 (1986).
8. Mano, N., Kanazawa, I. & Yamamoto, K. Voluntary movements and complex-spike discharges of cerebellar Purkinje cells. *Exp. Brain Res.* **17**, 265–280 (1989).
9. Bauswein, E., Kolb, F. P., Leimbach, B. & Rubia, F. J. Simple and complex spike activity of cerebellar Purkinje cells during active and passive movements in the awake monkey. *J. Physiol.* **339**, 379–394 (1983).
10. Wang, J. J., Kim, J. H. & Ebner, T. J. Climbing fiber afferent modulation during a visually guided multi-joint arm movement in the monkey. *Brain Res.* **410**, 323–329 (1987).

11. Ojakangas, C. L. & Ebner, T. J. Purkinje cell complex and simple spike changes during a voluntary arm movement learning task in the monkey. *J. Neurophysiol.* **68**, 2222–2236 (1992).
12. Ojakangas, C. L. & Ebner, T. J. Purkinje cell complex spike activity during voluntary motor learning: relationship to kinematics. *J. Neurophysiol.* **72**, 2617–2630 (1994).
13. Fu, Q. G., Mason, C. R., Flament, D., Coltz, J. D. & Ebner, T. J. Movement kinematics encoded in complex spike discharge of primate cerebellar Purkinje cells. *Neuroreport* **8**, 523–529 (1997).
14. Ito, M. *The Cerebellum and Neural Control* (Raven, New York, 1984).
15. Kawato, M. & Gomi, H. A computational model of four regions of the cerebellum based on feedback-error learning. *Biol. Cybern.* **68**, 95–103 (1992).
16. Houk, J. C., Buckingham, J. T. & Barto, A. G. Models of the cerebellum and motor learning. *Behav. Brain Sci.* **19**, 368–383 (1996).
17. Arbib, M. A. Spanning the levels in cerebellar function. *Behav. Brain Sci.* **19**, 434–435 (1996).
18. Kitazawa, S., Kohno, T. & Uka, T. Effects of delayed visual information on the rate and amount of prism adaptation in the human. *J. Neurosci.* **15**, 7644–7652 (1995).
19. Abeles, M. & Goldstein, M. H. Multiple train analysis. *Proc. IEEE* **65**, 762–773 (1977).

Acknowledgements. We thank F. A. Miles and K. Kawano for comments on the manuscript; M. Okui for technical assistance; K. Kurata for useful advice; and S. Yamane and R. Suzuki for encouragement during this study.

Correspondence and requests for materials should be addressed to S.K. (e-mail: kitazawa@etl.go.jp).

Synaptic vesicles retain their identity through the endocytic cycle

Venkatesh N. Murthy & Charles F. Stevens

Howard Hughes Medical Institute and Molecular Neurobiology Laboratory, Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

After fusion of synaptic vesicles with presynaptic membrane and secretion of the contents of the vesicles into the synaptic cleft (a process known as exocytosis), the vesicular membrane is retrieved by endocytosis (internalization) for re-use^{1,2}. Several issues regarding endocytosis at central synapses are unresolved, including the location of membrane retrieval (relative to the active zone, where exocytosis occurs), the time course of various endocytic steps, and the recycling path taken by newly endocytosed membranes. The classical model of synaptic-vesicle recycling, proposed by analogy to other cellular endocytic pathways, involves retrieval of the membrane, fusion of the membrane with endosome-like compartments and, finally, budding of new synaptic vesicles from endosomes¹, although the endosomal station may not be obligatory³. Here we test the classical model by using the fluorescent membrane dye FM1-43 (refs 4–6) with quantitative fluorescence microscopy. We find that the amount of dye per vesicle taken up by endocytosis equals the amount of dye a vesicle releases on exocytosis; therefore, we conclude that the internalized vesicles do not, as the classical picture suggests, communicate with intermediate endosome-like compartments during the recycling process.

When vesicular membrane internalization occurs in the presence of FM1-43, the dye is trapped inside the endocytosed vesicles^{4,5}. Then, when the dye is removed from the external medium, the FM1-43-labelled vesicles can be detected because of their fluorescence. After mild synaptic activity in the presence of FM1-43, the histogram of fluorescence intensity (which reflects the amount of membrane retrieved^{4,5}) shows equally spaced peaks^{7,8}. Such peaks revealed quantized endocytosis. In the present experiments we aimed to load a few vesicles per synapse with FM1-43 and to measure the decrease in fluorescence when some of these stained vesicles were released. If vesicles maintain their identity through the recycling process, then fluorescence should decrease with release in the same quantal steps as those seen with endocytosis (Fig. 1a, right). If, however, fusion with an endosomal compartment is part of the vesicle-recycling process, the dye should be shared with that compartment and the decrease in fluorescence seen when vesicles are released should occur with smaller steps (Fig. 1a, left).

We first measured the fluorescence change that corresponds to endocytosis of a single vesicle under the conditions of the present experiments. To stain synaptic vesicles, we evoked five action potentials at 1 Hz in the presence of 10 μM FM1-43 (10-s exposure). After washing in dye-free medium for 5 min, the field was imaged. Figure 1c shows a region of neuronal processes imaged simultaneously with differential-interference contrast (DIC) and fluorescence optics. Bright spots corresponding to synaptic boutons can be seen across the field. These spots are of the same size and shape as subresolution fluorescence beads (0.17 μm in diameter) imaged using the same settings. Therefore, the source of the fluorescence at each punctum is probably a few stained vesicles that are close together—in other words, a single synaptic bouton. When the same field was imaged after evoking 1,000 action potentials at 20 Hz in the absence of dye, most of the spots became dimmer. This loss of fluorescence was due to exocytosis of stained synaptic vesicles, as it does not occur in the absence of action potentials and occurs more

slowly in low external Ca^{2+} . In some experiments, a second round of staining with stronger stimuli (100 initial action potentials) was used to confirm the existence of functional synapses at these spots.

The distribution of the releasable fluorescence (ΔF), calculated as the difference in the intensity of the spots before and after the destaining stimulus of 1,000 action potentials, is shown for a single experiment in Fig. 2a, and for 452 boutons from nine experiments in Fig. 2b. The fluorescence-intensity distribution has peaks at integral multiples of 8.4 fluorescence units, which we interpret as corresponding to integral numbers of synaptic vesicles^{7,8}. Further experiments support this conclusion. First, altering the concentration of FM1-43 from 10 μM to 15 μM shifted the peak spacing by a factor of 1.43, which is within 5% of the expected value of 1.5. Second, loading with a different number of action potentials altered the relative sizes of the peaks, but not the peak spacing. Third, the fluorescence expected from a membrane area corresponding to that of a single vesicle, estimated by staining surface membranes⁷, is within 20% of the peak spacing (8.4 ± 0.8 units for peak spacing, versus 10.2 ± 0.7 units for surface-membrane fluorescence; numbers are mean $\pm$ standard deviation). Finally, and most important, the entire intensity distribution could be predicted quantitatively from the release-probability distribution estimated previously⁷, as follows. The distribution of releasable fluorescence was fitted by a sum of gaussian distributions with equally spaced peaks. The variance in each peak was assumed to have two sources, namely, measurement variance, which was estimated from acquiring repeated images of the same stained boutons, and variance of synaptic vesicle sizes, which was estimated from ultrastructural data⁹. The area under each gaussian curve, corresponding to the

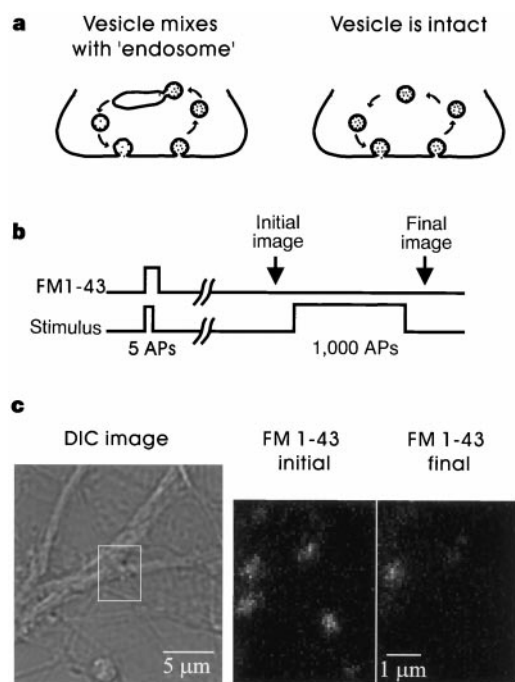


Figure 1 Experimental protocol for resolving single-vesicle fluorescence after staining synapses with FM1-43. **a**, Two models of recycling of newly endocytosed vesicles. In the first model (left), endocytosed vesicles with FM1-43 molecules (small asterisks) trapped inside mix with an endosomal compartment. When a new vesicle is generated, it will have fewer FM1-43 molecules. The model on the right assumes that the newly endocytosed vesicle remains intact, and does not mix with an endosomal compartment, as it is recycled. **b**, Experimental procedure used. We aimed to label a few vesicles to allow reliable detection of peaks in the intensity distribution. Field electrodes were used to stimulate action potentials in all the neurons within a region of the coverslip, thus increasing the number of stimulated synapses in the visualized field. The duration for which FM1-43 was present after stimulation was reduced, decreasing background staining. Five action potentials (APs) were elicited at 1 Hz in the presence of FM1-43. The preparation was washed in dye-free medium for 5 min, and two images (initial images) were acquired. After applying a destaining stimulus of 1,000 action potentials at 20 Hz to evoke release of all the stained vesicles, two more images (final images) were obtained. **c**, DIC image of a field of neural processes is shown at the left. The right-hand panels show fluorescence views of the area enclosed by the rectangle after loading with FM1-43 (FM1-43 initial) and after destaining with 1,000 action potentials (FM1-43 final). Most of the bright spots seen after FM1-43 loading have become dimmer after destaining stimuli (one exception is seen to the left of the frame).

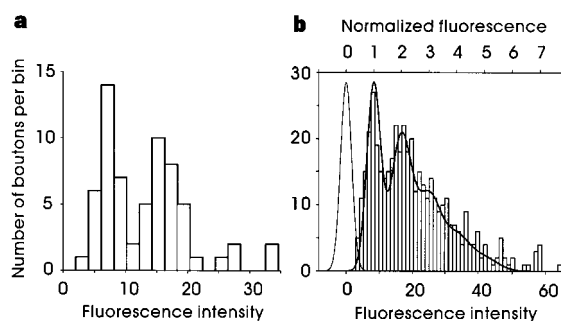


Figure 2 Distribution of fluorescence intensity at synaptic boutons after stimulation with five action potentials. **a**, Distribution of releasable fluorescence for 64 boutons from one experiment. The average fluorescence intensity in a 7×7 box centred over each bouton was calculated before and after destaining. The distribution shows two peaks centred on integral multiples of 8.3 units. **b**, Distribution of fluorescence for 452 boutons from nine experiments (histograms). Raw fluorescence intensities and fluorescence intensities normalized to the peak spacing of 8.4 units are shown on the top and bottom axes, respectively. The thick curve is the best fitting sum of gaussian distributions with peaks and variances, given by equation (3) (see Methods). The thin line centred on the origin is a gaussian curve with variance equal to the measurement variance, estimated from repeated images of the same field.

total number of observations for that gaussian, was obtained from the release-probability distribution (see Methods). Therefore, the entire distribution could be estimated with just one free parameter, namely the peak spacing. The resulting fit, shown by the continuous line in Fig. 2b, is good.

We used the estimate of single-vesicle fluorescence to determine whether the number of vesicles endocytosed accurately reflects the number of vesicles released. In response to a stimulus of 50 action potentials evoked at 20 Hz, the synaptic current recorded in a single neuron rapidly decreases in size during the first few stimuli, and then reaches a relatively steady response which is a small fraction of the initial response (Fig. 3a); this decline in response amplitude reflects depletion of the readily releasable pool of vesicles^{10–12}. Presynaptic terminals were stained with FM1-43 using a stimulation protocol of 50 action potentials at 20 Hz, a procedure which empties the readily releasable pool but causes little or no release of vesicles located outside the pool^{11,12}. After destaining with a long stimulus train, ΔF was estimated for 247 boutons in three experiments (Fig. 3b). The distribution of ΔF , converted to number of vesicles using the value for single-vesicle fluorescence, matches the distribution of the number of docked vesicles measured ultrastructurally⁹. We found the average number of vesicles in the readily releasable pool to be 7.7 ± 4.4 , which closely concurs with the average size of readily releasable pools estimated electrophysiologically^{10,12} and anatomically⁹. These results indicate that the population of FM1-43-labelled vesicles comprises most, if not all, of the endocytosed vesicles.

We next tested the hypothesis that synaptic vesicles, once endocytosed, retain their identity through the rest of the recycling

pathway until they exocytose again. While the vesicles remain intact, FM1-43 molecules that are trapped in the vesicle will be unable to mix with other cellular compartments, and therefore will not be diluted as the vesicle moves through the recycling pathway. Then, when exocytosis occurs, the amount of dye released per vesicle will be equal to the amount of dye taken up during endocytosis. To test this hypothesis, one could in principle compare fluorescence before and after a single action potential across a population of synapses. However, as the probability of releasing a stained vesicle by a single action potential is <0.005 (estimated from the average release probability and the fraction of vesicles that are stained), in practice many action potentials are necessary to release one stained vesicle with high probability. Therefore, we used a releasing stimulus of 100 action potentials at 20 Hz, which quickly releases the readily releasable pool of vesicles, but releases only a few vesicles outside this pool (Fig. 3). In effect, then, we are determining how many of the stained vesicles, produced by five action potentials, entered the readily releasable pool.

In six experiments, after staining synaptic boutons with FM1-43 using a stimulus of five action potentials and obtaining two images, we did a partial destaining with 100 stimuli at 20 Hz. Two images were taken 20 s after ending the stimulus, to ensure that the dye partitions out of the membrane (dye off-rate was measured at <3 s). Two final images were taken after complete destaining with 1,000 stimuli at 20 Hz. The total releasable fluorescence and the amount of fluorescence released as a result of the partial destaining stimulus were estimated. The distribution of partial fluorescence loss (ΔF_p) for 329 boutons is shown in Fig. 4a. Many of the boutons did not release any fluorescence during the partial destaining stimulus, and

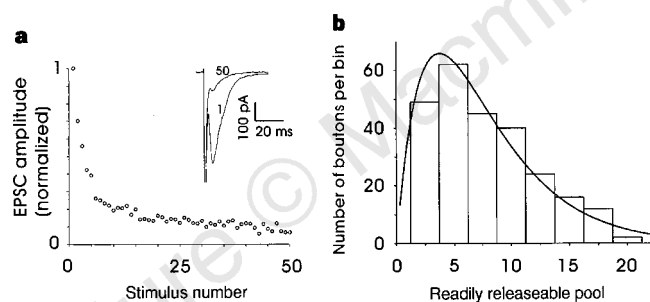


Figure 3 Estimate of the readily releasable pool of vesicles. **a**, Whole-cell autaptic currents were recorded from a single neuron grown in a microisland. In response to a train of 50 action potentials at 20 Hz, the peak excitatory postsynaptic current (EPSC), normalized to the first response, shows rapid depression. Inset shows the first and the 50th responses. The initial artefact arises from the action-potential stimulation. **b**, Synaptic boutons were stained with FM1-43 by stimulating them at 20 Hz for 2.5 s (total of 50 action potentials). This stimulus protocol was designed to release the readily releasable pool of synaptic vesicles, with negligible refilling of this pool. FM1-43-containing solution was perfused for an extra 30 s to maximize this loading of dye into vesicles. The total amount of releasable fluorescence was determined by acquiring images before and after stimulating the synapses with 1,000 action potentials. Dividing the fluorescence intensity by the value for single-vesicle fluorescence (8.4 units) provided an estimate for the number of vesicles released (and recovered by endocytosis). The histogram shows the distribution of the readily releasable pool for 248 boutons in three experiments. The average size of the readily releasable pool was 7.7 ± 4.4 vesicles. The continuous line is the fit to $47.5Ne^{-0.264N}$, where N is the readily releasable pool.

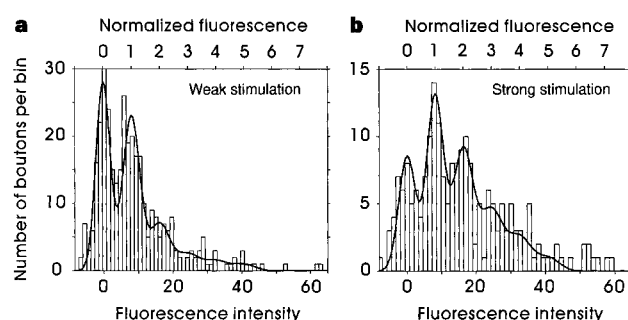


Figure 4 Fluorescence loss occurs in a quantized manner. **a**, Distribution of fluorescence lost after partial destaining. Synaptic terminals were stained with FM1-43 during stimulation with five action potentials. After acquiring two images, 100 stimuli were delivered at 20 Hz. Two more images were acquired before 1,000 stimuli were delivered at 20 Hz; two final images were then acquired. The fluorescence difference between initial and final images gave the total releasable fluorescence, which is the same as that in Fig. 2. The difference between initial and middle images (after 100 action potentials) gives an estimate of the fluorescence loss due to the loss of a small number of stained vesicles. This distribution of raw and normalized fluorescence for 329 boutons (bars) shows a prominent peak at 8.2 units, corresponding to a single vesicle's fluorescence. The continuous line is the best fitting sum of gaussian distributions, with variances given by equation (1). The peak at the origin is composed of boutons that did not destain during the 100 action-potential stimuli (and did destain during the subsequent 1,000 stimuli). The variances of the gaussians are larger than those of the corresponding peaks for the total ΔF (Fig. 2b), which is as expected because of the larger remaining fluorescence after partial destaining compared with after complete destaining. **b**, Distribution of fluorescence lost after partial destaining when synaptic terminals were loaded with dye using stronger stimulation. Synaptic boutons were stimulated at 20 Hz for 5 s and, after a 20-s delay, FM1-43 solution was perfused for 10 s. The rest of the experiment was similar to the one in **a**. The distribution of fluorescence loss due to the partial destaining stimulus once again exhibits a prominent peak at 8.2, a value corresponding to the single-vesicle fluorescence.

formed a peak centred at the origin. The distribution of fluorescence for those boutons that did lose some fluorescence shows a prominent peak in fluorescence intensity at 8.2 units, nearly matching the peak spacing obtained for FM1-43 loading (8.4 units; Fig. 2). The difference between the two values is within the experiment-to-experiment variability of peak spacing, which was $\sim 8\%$. Therefore, when a vesicle is released, the bouton loses an amount of dye that is equal to the amount trapped by a single vesicle when it was endocytosed.

We interpret the destaining fluorescence quantum as corresponding to single vesicles, but it is possible that, in fact, multiple vesicles shared dye during the recycling pathway, and that all were then released in the 100-action-potential stimulus. This possibility is, however, unlikely on statistical grounds. For two vesicles that shared a single quantum of fluorescence during recycling, the probability that both would be released at most boutons is $<10^{-4}$, assuming that stained vesicles are distributed homogeneously within the population of synaptic vesicles⁶; the probability is even smaller if more vesicles share dye. It is also possible that we somehow saturated some intermediate endosome-like compartment with FM1-43, so that vesicles generated from this compartment had a constant amount of dye. This seems unlikely because, across different synapses, a different number of vesicles will be endocytosed, so different total amounts of dye will have to be shared with an intermediate compartment. As we always see sharp peaks during release, there must be little sharing of dye between vesicles.

In the above experiments, dye loading was achieved by a very mild stimulus of just five action potentials. Perhaps the recycling pathway only uses an endosomal compartment during strong stimulation of the synapses. To test whether endocytosed vesicles retain their identity after stronger synaptic activity, we delivered a stimulus train of 100 action potentials at 20 Hz to the neurons. After a delay of 20 s after the end of the stimulus train, neuronal processes were superfused with FM1-43-containing solution for 10 s. The delay between the end of stimulus and onset of superfusion of FM1-43 was needed to ensure labelling of only a few vesicles. If the dye was superfused immediately after stimulation, the resulting uptake of dye corresponded to many vesicles and peaks in the intensity distribution could not be resolved because of the increased variances that accompany the labelling of more than about four vesicles.

After staining boutons, partial and complete destaining stimuli were applied as above. Figure 4b shows the distribution of partial fluorescence loss, and peaks are again seen at locations corresponding to integer multiples of single-vesicle fluorescence. Therefore, even after strong stimulation, endocytosis occurs in a quantal fashion and the vesicles retain their identities as they move through the recycling pathway and become fusion-competent. The existence and similarity of peaks for early- and late-dye application also indicates that the overall rate of endocytosis, which has a time constant of 20–30 s (ref. 6), is defined by the number of vesicles endocytosed per unit time, rather than by the rate of the endocytic processes for individual vesicles.

Much of our previous knowledge about endocytosis has come from ultrastructural observations. At the neuromuscular junction, following intense stimulation, omega-shaped structures and bulk invaginations of the membrane can be seen in the presynaptic membrane. On the basis of these observations, the classical model of recycling, involving membrane internalization and mixing with endosomal structures, was proposed¹. Whether this model applies to small central synapses under more physiological stimulation has, however, been unknown. Several studies that have selectively disrupted proteins, including dynamin and amphiphysin, that are important for endocytosis have found that endocytosis is arrested at the stage of clathrin-coated-pit formation^{13–15}. These coated pits, which are uniform in size, may represent the mechanism for quantized endocytosis. Our results are compatible with a model

of vesicle recycling in which endocytosis occurs mainly through internalization of vesicle-sized membrane patches that are exposed to the extracellular space and which then remain intact throughout the vesicle-recycling pathway¹⁶. Whether the classical recycling pathway involving endosomes can be recruited at these small central synapses under some other conditions is unknown. □

Methods

Culture preparation and electrophysiology. Hippocampal neurons were cultured using described methods^{7,17}. Cultures were allowed to mature for 10–15 days before experiments were conducted. Whole-cell autaptic currents were recorded from isolated single neurons grown in microislands¹¹. The patch pipettes (2–3 M Ω) contained (in mM): 120 potassium gluconate, 10 KCl, 5 Mg-ATP, 0.3 GTP, 10 HEPES and 0.2 EGTA. Cells were perfused with extracellular medium containing (in mM): 136 NaCl, 2.5 KCl, 10 glucose, 10 HEPES, 2 CaCl₂ and 1.3 MgCl₂, with 50 μ M DL-APV (Research Biochemicals).

FM1-43 loading and destaining. For the imaging experiments, a plexiglass chamber was mounted on a movable stage of a microscope. The chamber had a circular aperture. The coverslip with neurons was directly attached to the rim of the aperture with vacuum sealant (Dow Corning). The coverslip formed the bottom of the chamber, and neurons could be viewed with an oil lens (magnification $\times 40$, 1.0 NA, Fluotar, Leica) of the inverted microscope. The extracellular medium contained 10 μ M DNQX and 50 μ M DL-APV to block recurrent activity. Two silver chloride wires, separated by ~ 5 mm and located close to the surface of the coverslip, were used to evoke action potentials. A constant current source was used to pass 10 mA for 1 ms; this was above the threshold for eliciting action potentials. Whole-cell recordings confirmed that the generation of action potentials could follow stimulation rates of up to 20 Hz for as long as 3 min. Synaptic vesicles were loaded with FM1-43 by bathing the cultures in medium containing 10 μ M dye, and eliciting a variable number of action potentials in the neurons at 0.5 or 1 Hz. The dye was applied through a puffer pipette with a tip diameter of ~ 2 μ m. Control experiments ensured that the dye concentration was uniform within the field of view. The total time of dye application was usually 15 s (it was 10 s in some later experiments), and was followed by >5 min perfusion with dye-free medium. Further stimulation with action potentials caused release of dye-containing vesicles; this was seen as a loss of fluorescence. Several destaining methods were used (described above).

Imaging and analysis. A Leica TCS NT confocal laser-scanning system attached to a Leica DM IRBE inverted microscope was used for imaging. The 488-nm line of an argon-ion laser was used for excitation, and the emitted light was filtered through a 510-nm long-pass filter cube and detected by a photomultiplier. A second photomultiplier was used to obtain transmitted light images of the preparation. Images were 512 $\times$ 512 pixels wide, with a pixel width of 0.081 μ m (total image dimension 41.6 μ m $\times$ 41.6 μ m). Each image was an average of two successive frames, and two such images were obtained for each condition. The pinhole was set to the fully open position to obtain the greatest depth of field. The point-spread function in the xy plane had a half-maximal width of ~ 0.5 μ m.

Images were analysed using custom-written routines in MATLAB. Regions of interest (ROIs) around isolated bright spots were identified, and the centre of mass of each spot was calculated. The average intensity over a 7×7 pixel box around the centre of mass was calculated. Fluorescence was expressed in intensity units that correspond to fluorescence values averaged over all pixels within the 7×7 box (fluorescence value for each pixel varied from 0 to 255). This procedure was repeated for each frame. As ROIs were defined by rectangles, they usually included a small background area. However, the contribution of the background would be removed as the fluorescence value after complete destaining was subtracted from all frames. In a few experiments there were slight shifts in successive images. To account for these, the region of interest was shifted to be centred around the centre-of-brightness of the fluorescent spot⁶. A few experiments in which excessive shifts in images occurred, presumably because of changes in focal distance, were discarded.

The observed distribution of fluorescence was fitted to a sum of gaussian distributions centred at integral multiples of the peak spacing; this parameter was estimated by the method of least squares (see below). The variance for each gaussian is the sum of two variances—measurement variance and variance due to variability in diameters of synaptic vesicles. We estimated the measurement

variance from repeated images of the puncta in the absence of stimulation. The coefficient of variation of measurement, c_m , was found to be roughly constant (0.05) as the mean intensity increased. As each ΔF measurement was a difference between two images (initial and final), the variance is the sum of the variance for each image. The residual fluorescence was typically <10 units (which is comparable to fluorescence of a single vesicle), and was similar across different boutons. Therefore, the measurement variance for successive peaks is given by:

$$c_m^2((\mu_k + r)^2 + r^2)$$

where μ_k is the centre of the k th peak and r is the residual fluorescence. The variance in fluorescence due to vesicle-size variability is given by:

$$c_v^2 \mu_k^2 k$$

where c_v is the coefficient of variation of vesicle surface area (0.2, estimated from ultrastructure⁹). Therefore, the total variance for the k th peak is:

$$\sigma_k^2 = c_m^2((\mu_k + r)^2 + r^2) + c_v^2 \mu_k^2 k \quad (1)$$

The height of the peaks can be obtained from the release probability distribution $w(p)$. For a given synapse with release probability p , the distribution of the number of releases is given by a binomial distribution. For a population of synapses with a release probability distribution $w(p)$, this distribution (for five stimuli) becomes:

$$g(k) = \int_0^1 dp \binom{5}{k} p^k (1-p)^{5-k} w(p) \quad (2)$$

The release-probability distribution $w(p)$ has been measured for these synapses⁷:

$$w(p) = \frac{\lambda^2}{2} p e^{-\lambda p} \quad \lambda = 8;$$

Therefore, $g(k)$ can be determined exactly, and will give the areas under each of the peaks in the ΔF distribution. Knowing the height and the variance of each peak (as a function of the mean), we can fit the staining distribution with only one free parameter, μ_k .

$$P(f) = \sum_{k=1}^5 g(k) e^{-\frac{1}{2} \left(\frac{f - \mu_k}{\sigma_k} \right)^2} \quad (3)$$

Maximum-likelihood fits gave a peak spacing of 8.4 fluorescence units for the staining histogram (Fig. 2b) and 8.2 fluorescence units for the partial destain histogram (Fig. 4a, b).

Received 19 September; accepted 5 January 1997.

- Heuser, J. E. & Reese, T. S. Evidence for recycling of synaptic vesicle membrane during transmitter release at the frog neuromuscular junction. *J. Cell Biol.* **57**, 315–344 (1973).
- Ceccarelli, B., Hurlbut, W. P. & Mauro, A. Turnover of transmitter and synaptic vesicles at the frog neuromuscular junction. *J. Cell Biol.* **57**, 499–524 (1973).
- Takei, K., Mundigl, O., Daniell, L. & De Camilli, P. The synaptic vesicle cycle: a single vesicle budding step involving clathrin and dynamin. *J. Cell Biol.* **133**, 1237–1250 (1996).
- Betz, W. J. & Bewick, G. S. Optical analysis of synaptic vesicle recycling at the frog neuromuscular junction. *Science* **255**, 200–203 (1992).
- Betz, W. J., Mao, F. & Bewick, W. J. Activity dependent staining and destaining of living motor nerve terminals. *J. Neurosci.* **12**, 363–375 (1992).
- Ryan, T. A. & Smith, S. J. Vesicle pool mobilization during action potential firing at hippocampal synapses. *Neuron* **14**, 983–989 (1995).
- Murthy, V. N., Sejnowski, T. J. & Stevens, C. F. Heterogeneous release properties of visualized individual hippocampal synapses. *Neuron* **18**, 599–612 (1997).
- Ryan, T. A., Reuter, H. & Smith, S. J. Optical detection of quantal presynaptic membrane turnover. *Nature* **388**, 478–482 (1997).
- Schikorski, T. & Stevens, C. F. Quantitative ultrastructural analysis of hippocampal excitatory synapses. *J. Neurosci.* **17**, 5858–5867 (1997).
- Stevens, C. F. & Tsujimoto, T. Estimates for the pool size of releasable quanta at a single central synapse and for the time required to refill a pool. *Proc. Natl Acad. Sci. USA* **92**, 846–849 (1995).
- Rosenmund, C. & Stevens, C. F. Definition of the readily releasable pool of vesicles at hippocampal synapses. *Neuron* **16**, 1197–1207 (1996).
- Dobrunz, L. E. & Stevens, C. F. Heterogeneity of release probability, facilitation, and depletion at central synapses. *Neuron* **18**, 995–1008 (1997).
- Takei, K., McPherson, P. S., Schmid, S. L. & De Camilli, P. Tubular membrane invaginations coated by dynamin rings are induced by GTP- γ S in nerve terminals. *Nature* **374**, 186–190 (1995).
- Hinshaw, J. E. & Schmid, S. L. Dynamin self-assembles into rings suggesting a mechanism for coated vesicle budding. *Nature* **374**, 190–192 (1995).
- Shupliakov, O. *et al.* Synaptic vesicle endocytosis impaired by disruption of dynamin-SH3 domain interactions. *Science* **276**, 259–263 (1997).
- Cremona, O. & De Camilli, P. Synaptic vesicle endocytosis. *Curr. Opin. Neurobiol.* **7**, 323–330 (1997).
- Bekkers, J. M. & Stevens, C. F. NMDA and non-NMDA receptors are co-localized at individual excitatory synapses in cultured rat hippocampus. *Nature* **341**, 230–233 (1989).

Acknowledgements. We thank C. Boyer for preparation of cultures. V.N.M. is a Helen Hay Whitney Foundation fellow. This research was supported by grants from the NIH and HHMI to C.F.S.

Correspondence and request for materials should be addressed to C.F.S. (e-mail: cfs@salk.edu).

C. elegans phagocytosis and cell-migration protein CED-5 is similar to human DOCK180

Yi-Chun Wu & H. Robert Horvitz

Howard Hughes Medical Institute, Department of Biology, 68-425, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

During programmed cell death, cell corpses are rapidly engulfed¹. This engulfment process involves the recognition and subsequent phagocytosis of cell corpses by engulfing cells^{1–4}. How cell corpses are engulfed is largely unknown. Here we report that *ced-5*, a gene that is required for cell-corpse engulfment in the nematode *Caenorhabditis elegans*⁵, encodes a protein that is similar to the human protein DOCK180 and the *Drosophila melanogaster* protein Myoblast City (MBC), both of which have been implicated in the extension of cell surfaces⁶. *ced-5* mutants are defective not only in the engulfment of cell corpses but also in the migrations of two specific gonadal cells, the distal tip cells. The expression of human DOCK180 in *C. elegans* rescued the cell-migration defect of a *ced-5* mutant. We present evidence that *ced-5* functions in engulfing cells during the engulfment of cell corpses. We suggest that *ced-5* acts in the extension of the surface of an engulfing cell around a dying cell during programmed cell death. We name this new family of proteins that function in the extension of cell surfaces the CDM (for CED-5, DOCK180 and MBC) family.

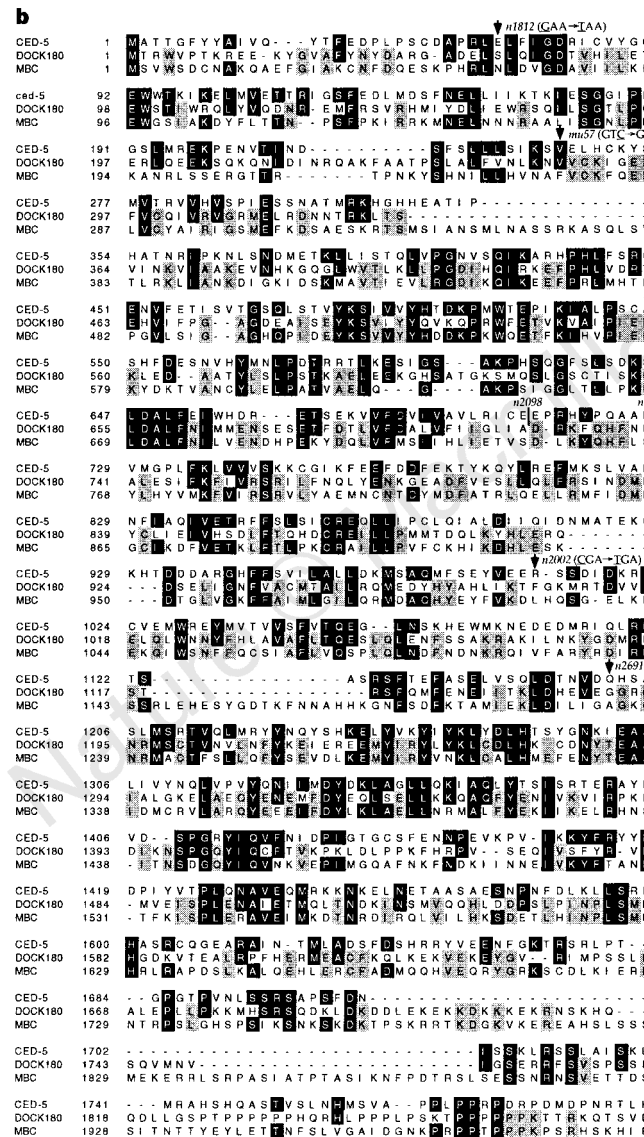
In *C. elegans*, the engulfment of cell corpses is regulated by at least six genes, *ced-1*, *ced-2*, *ced-5*, *ced-6*, *ced-7* and *ced-10* (*ced*, for cell death abnormal)^{5,7}. Mutations in any of these genes block the engulfment of many cell corpses and hence cause a mutant phenotype that is characterized by the persistence of cell corpses. Ultrastructural studies revealed that these mutations prevent extension of the membranes of the engulfing cells around the dying cells^{5,7}. We observed that *ced-2*, *ced-5* and *ced-10* mutants, but not *ced-1*, *ced-6* or *ced-7* mutants, are also defective in the specific migration of the gonadal distal tip cells (DTCs). This defect has been independently observed by K. Nishiwaki and M. Hengartner (personal communications). The DTCs are located at the tips of the two gonadal arms and guide the extension of each growing gonadal arm during larval development^{8,9}. In *ced-2*, *ced-5* and *ced-10* mutants, the DTCs frequently make extra turns or stop prematurely, which results in abnormally shaped gonads. For example, 77 per cent of *ced-5(n1812)* mutant animals ($n = 120$) showed abnormal DTC migration.

To understand how CED-5 functions in both the engulfment of cell corpses and the migration of the DTCs, we cloned the gene. *ced-5* maps between *mec-3* and *him-8* on chromosome IV (ref. 5). These two genes define an approximately 100-kilobase (kb) region on the physical map^{10,11} (P. Meneely, personal communication; Fig. 1a). We used genomic DNA clones from this interval to rescue the *Ced-5* mutant phenotype of persisting cell corpses in germline transformation

Table 1 Expression of DOCK180 rescues the DTC-migration defect in *ced-5* mutants

hsp transgene*	<i>ced-5</i>	DOCK180	GFP	none
Animals with DTC-migration defect (%)	8 $\pm$ 2	29 $\pm$ 5	69 $\pm$ 4	72 $\pm$ 5

* The heat-shock constructs were injected into *ced-5(n1812); unc-76(e911)* animals (see Methods). All data were obtained after heat shock. The DTC-migration defect was scored on the basis of the shape of the gonad in early adults. Wild-type animals ($n = 181$) show no defect in DTC migration following heat shock. Data are means $\pm$ s.e.m. from at least two stably transmitting lines. At least 80 animals were scored from each line. GFP, green fluorescent protein¹⁷.



experiments and localized the rescuing activity to a 10.6-kb genomic DNA fragment (Fig. 1a). We isolated *ced-5* complementary DNAs using this fragment and defined the 5' end of the *ced-5* messenger RNA using the RACE (for rapid amplification of cDNA ends) method. The sequences of these cDNAs revealed a 1,781-amino-acid open reading frame, an SL1 *trans*-spliced leader sequence found at the 5' end of many *C. elegans* transcripts¹², and a 3'-end poly(A) tract, indicating that we had identified a complete *ced-5* transcription unit. Northern blot analysis, using a *ced-5* cDNA as probe, showed a single band of 5.6 kb (data not shown), consistent with the size of the full-length *ced-5* cDNA. The expression of a *ced-5* DNA in *ced-5(n1812)* animals under the control of *C. elegans* heat-shock promoters (*hsp*)¹³ rescued the defects in both DTC migration (Table 1) and cell-corpse engulfment (Table 2; see below), indicating that the *ced-5* cDNA encodes a functional

NATURE | VOL 392 | 2 APRIL 1998

CED-5 protein. We identified molecular lesions in six *ced-5* mutant alleles: three, *n1812*, *n2002* and *n2691*, have nonsense mutations (at codon positions 28, 962 and 1,145, respectively); one allele, *mu57*, has a single-base deletion (in codon 216); two other alleles, *n2098* and *n2099*, have single-base changes (in the splice-acceptor sequences of the sixth and eighth introns, respectively) (Fig. 1b). The allele *ced-5(n1812)* is likely to be null as this premature stop codon (UAA) eliminates more than 98% of the CED-5 protein and the engulfment defect of *n1812* homozygotes is indistinguishable from that of *n1812/sDf2* heterozygous animals⁵ (*sDf2* is a deletion that spans the *ced-5* locus).

The 1,781-amino-acid sequence of CED-5 is most similar to the sequence of the human protein DOCK180. These proteins share 26% identity over their entire lengths (Fig. 1b). CED-5 and DOCK180 both share significant sequence similarity with the amino-acid sequences predicted from the *Drosophila* gene *mbc*¹⁴ (Fig. 1b), the human cDNA clone KIAA0209 (ref. 15), the yeast open reading frame L9576.7 (GenBank accession number 664878) and a mouse expressed sequence tag (EST) (GenBank accession number AA110899).

To determine whether human DOCK180 and CED-5 are functionally related as well, we tested the ability of *hsp::DOCK180* to rescue the *Ced-5* mutant phenotype. Inducing the expression of DOCK180 rescued the DTC-migration defect (Table 1) but not the corpse-engulfment defect (data not shown). This partial rescue by DOCK180 of the *Ced-5* mutant phenotype could indicate either a need for a higher level of gene activity for cell-corpse engulfment than for DTC migration or a bifunctional role for CED-5, such that DOCK180 can supply the function needed for DTC migration but not that for corpse engulfment. As DOCK180 rescued the CED-5 defect in DTC migration, DOCK180 and CED-5 are at least in this

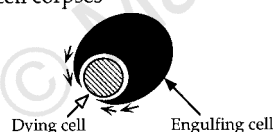
respect functionally interchangeable.

DOCK180 was isolated on the basis of its interaction with the cytoskeleton-associated adaptor protein CRK⁶, which consists mainly of SH2 and SH3 domains¹⁶ and has been implicated in integrin-mediated signalling and cell movement¹⁷. Vitronectin, a member of the integrin superfamily, has been implicated in the engulfment of cell corpses in mammals¹⁸. The expression of DOCK180 in 3T3 fibroblasts can cause these cells to extend their surfaces and adopt flat and polygonal shapes, indicating that DOCK180 can regulate the extension of cell surfaces⁶. *Drosophila* MBC is necessary for myoblast fusion and for the migrations of some epithelial cells, both of which require the extension of cell surfaces, presumably through reorganization of the cytoskeleton^{14,19}. We propose that CED-5, DOCK180 and MBC all function to mediate the extension of cell surfaces and that CED-5 does so by acting in engulfing cells during the phagocytosis of cell corpses as well as in migrating DTCs. *ced-5* mutants appear to be normal in the migrations of other cells (for example, the P1–P12 precursor cells and the HSN neurons) as well as in axonal outgrowth and in cell fusions involved in the development of the hypodermal syncytium (ref. 5; and data not shown). Thus, CED-5 is likely to be involved in a specific type of membrane extension that is common to cell-corpse engulfment and DTC migration.

As already noted, we rescued the engulfment defect of a *ced-5* mutant using heat shock to induce the expression of a wild-type *ced-5* transgene (Table 2). We believe that this rescue was effected by the expression of *ced-5* in engulfing cells rather than in cell corpses, for two reasons. First, the defect in the mutant animals subjected to heat-shock was rescued even at late stages of larval development. The finding that expression of *ced-5*, even at a time when cell corpses have presumably long been dead (for hours in L1 and L2 larval stages, or even days in L3 and L4 larval stages), still rescued the defect (Table 2; larvae column) suggests that *ced-5* functions not in cell corpses but rather in engulfing cells. Second, using the green fluorescent protein (GFP) as a reporter²⁰, we found that *ced-5(n1812)* animals carrying a GFP transgene under the control of these *C. elegans* heat-shock promoters did not express GFP in persisting cell corpses following heat shock, but they did express GFP in other somatic cells, including engulfing cells, throughout larval development (data not shown). This finding indicates that the transcriptional and/or translational machineries are probably inactive in persisting cell corpses. We conclude that rescue of the *ced-5* engulfment defect by the heat-shock-inducible *ced-5* transgene is caused by the expression of this transgene in engulfing cells.

We used a similar strategy to explore whether *ced-5* functions in engulfing cells during the removal of germline cell corpses in addition to its function in engulfing cells during the removal of somatic cell corpses. Germline cell corpses are engulfed by the gonadal sheath cells in a *ced-5*-dependent manner (M. Hengartner, E. Hartwig and H.R.H., unpublished observations). After heat shock of *ced-5* animals carrying an *hsp::GFP* transgene, we detected GFP expression in gonadal sheath cells but not in the germ line (data not shown). This finding is consistent with previous observations that these *C. elegans* heat-shock promoters do not drive the expression of transgenes in the germline²¹ (A. Fire, personal communication). *ced-5* animals carrying an *hsp::ced-5* transgene

Engulfment of cell corpses



DTC migration

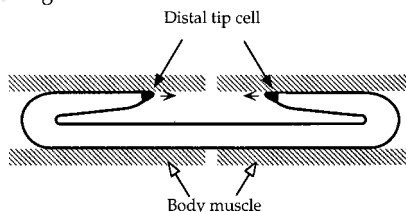


Figure 2 Cell-corpse engulfment and DTC migration are similar processes. In each case, the surface membrane of a cell (black) extends along the surface of another cell (hatched). The small arrows near the black cells indicate the directions of cell-surface extension. Only the relevant parts of body muscles are shown.

Table 2 Expression of a *ced-5* cDNA rescues the engulfment defect of *ced-5* mutants

<i>hsp</i> transgene*	Heat shock	No. of cell corpses in head†					No. of cell corpses in germ line‡
		4-fold embryo	L1	L2	L3	L4	
<i>ced-5</i>	+	0.8 ± 2	6 ± 4	3 ± 2	3 ± 2	2 ± 2	4 ± 5
	–	34 ± 2	30 ± 5	27 ± 3	21 ± 4	15 ± 3	27 ± 4
GFP	+	33 ± 2	30 ± 3	28 ± 2	20 ± 2	16 ± 4	24 ± 6

* The heat-shock constructs were injected into *ced-5(n1812); unc-76(e911)* animals (see Methods). Mixed-staged transgenic progeny were subjected to heat shock (+) or left at 20 °C (–).

† Transgenic animals were scored for the number of head cell corpses, which were generated during embryogenesis, and for developmental stage 10 h after heat shock.

‡ Number of cell corpses in each gonadal arm of the transgenic animal was scored 24 h after heat shock (see Methods). Data shown are means ± s.e.m. from two independent stably transmitting lines. More than 20 animals were scored from each line. GFP, green fluorescent protein¹⁷.

had significantly fewer germline cell corpses persisting after heat shock compared with heat-shocked *ced-5* animals carrying an *hsp::GFP* transgene (Table 2). This observation further suggests that *ced-5* functions in engulfing cells and not in persisting cell corpses during programmed cell death.

We have shown that *ced-5* is likely to act in engulfing cells during cell-corpus engulfment and also is required for the normal migration of DTCs. Both the engulfment of cell corpses^{5,22} and cell migration require a cell to extend its surface in a polarized fashion (Fig. 2), which is consistent with our hypothesis that CED-5, like DOCK180 and MBC, functions in cell-surface extension. On the basis of this consideration and our finding that DOCK180 rescued the DTC-migration defect of *ced-5* mutant animals, we propose that CED-5, DOCK180 and MBC define a new evolutionarily conserved gene family involved in the extension of cell surfaces. We call this family CDM. We propose that *ced-5* acts in phagocytosis in response to the recognition of dying cells during programmed cell death. By analogy to MBC, *ced-5* could mediate the cytoskeletal reorganization that occurs as an engulfing cell extends its cell surface around a dying cell. We suggest that, like other proteins involved in programmed cell death^{23–25}, CED-5 and the proteins with which it interacts during the engulfment of cell corpses are evolutionarily conserved. □

Methods

Germline transformation experiments. For genomic rescue experiments, DNA was co-injected into *ced-5(n1812)* animals at concentrations of 25–50 $\mu\text{g ml}^{-1}$ with the dominant roller marker pRF4 (50 $\mu\text{g ml}^{-1}$) as previously described²⁶. We scored persistent cell corpses in the head regions of L2 or L3 roller animals from stably transmitting transgenic lines using Nomarski optics as previously described²⁶. Non-rollers or non-rescued rollers had about 20–25 cell corpses. Roller animals containing 0–5 corpses were scored as rescued for the *ced-5* engulfment defect.

To construct *hsp::ced-5*, we did a three-component ligation reaction: we ligated the *KpnI*–*XhoI* fragment from the plasmid pC5OKBA, which contains the 5' half of the *ced-5* cDNA, with the *XhoI*–*Apal* fragment from the pC583 construct, which contains the 3' half of the *ced-5* cDNA, to the *KpnI*–*Apal* fragment from *hsp* vectors pPD49.78 or pPD49.83 (from A. Fire). To construct *hsp::GFP*, we excised the *XbaI*–*Apal* fragment from the plasmids pPD96.04 (from A. Fire) and Tu#61 (from M. Chalfie); these plasmids contain the GFP gene with and without a nuclear localization signal, respectively. We cloned the fragments into the vectors pPD49.78 and pPD49.83, previously digested with *NheI* and *Apal*. To make *hsp::DOCK180* constructs, we excised the *XhoI* fragment from plasmid pBIDOCK180, which contains a DOCK180 cDNA⁶, blunt-ended the fragment and cloned it into the vectors pPD49.78 and pPD49.83 through their *EcoRV* sites. The heat-shock constructs were co-injected into *ced-5(n1812);unc-76(e911)* animals at 50 $\mu\text{g ml}^{-1}$ with the *unc-76*-rescuing plasmid p76-16B²⁷ to establish transgenic lines and with the *egl-5::GFP* plasmid pSC212 to identify transgenic embryos (50 $\mu\text{g ml}^{-1}$ each) (A. Chisholm and H.R.H., unpublished results).

Heat-shock experiments. To determine the extent of rescue of the *ced-5* engulfment defect in somatic cell death, we heat-shocked mixed-staged transgenic animals for 1.5 h at 33 °C, followed by 10 h recovery at 20 °C. We counted persisting cell corpses in the heads of transgenic animals at different stages. To test for rescue of the engulfment defect in germline cell death, transgenic animals were given two pulses of heat shock over 36 h. In brief, we picked L4 transgenic animals and exposed them for 1.5 h to 33 °C. Following heat shock, animals that entered adulthood were selected and transferred to an incubator at 20 °C for recovery. After 10.5 h, we exposed these animals for 1.5 h to 33 °C. We counted germline cell corpses in each gonadal arm 22.5 h after this treatment. To test for rescue of the DTC-migration defect, L2 transgenic animals were selected and exposed to three 1.5-h heat-shock treatments at 33 °C, separated by two 10.5-h recovery intervals at 20 °C to span an approximately 25-h period during DTC migration. The DTC-migration

pattern was inferred from the gonadal morphology of young adults.

Characterization of *ced-5* genomic and cDNA structure. The 10.6-kb *Clal* genomic fragment containing *ced-5* rescuing activity was used to screen a mixed-staged cDNA library²⁸ and an embryonic cDNA library²⁹. We determined the sequences of *ced-5* genomic DNA and three overlapping *ced-5* cDNAs, using a Sequenase 2.0 kit (USB). We identified the 5' end of *ced-5* mRNA using a 5' RACE system (GIBCO-BRL) and identified mutations in *ced-5* alleles by determining the sequences of genomic regions produced by amplification with the polymerase chain reaction.

Received 15 October 1997; accepted 6 January 1998.

- Ellis, R. E., Yuan, J. & Horvitz, H. R. Mechanisms and functions of cell death. *Annu. Rev. Cell Biol.* **7**, 663–698 (1991).
- Savill, J., Fadok, V., Henson, P. & Haslett, C. Phagocyte recognition of cells undergoing apoptosis. *Immunol. Today* **14**, 131–136 (1993).
- Hart, S. P., Haslett, C. & Dransfield, I. Recognition of apoptotic cells by phagocytes. *Experientia* **52**, 950–956 (1996).
- Savill, J. Apoptosis in resolution of inflammation. *J. Leuk. Biol.* **61**, 375–380 (1997).
- Ellis, R., Jacobson, D. M. & Horvitz, H. R. Genes required for the engulfment of cell corpses during programmed cell death in *C. elegans*. *Genetics* **129**, 79–94 (1991).
- Hesegawa, H. *et al.* DOCK180, a major CRK-binding protein, alters cell morphology upon translocation to the cell membrane. *Mol. Cell. Biol.* **16**, 1770–1776 (1996).
- Hedgecock, E., Sulston, J. & Thomson, J. N. Mutations affecting programmed cell deaths in the nematode *Caenorhabditis elegans*. *Science* **220**, 1277–1279 (1983).
- Kimble, J. Alterations in cell lineage following laser ablation of cells in the somatic gonad of *Caenorhabditis elegans*. *Dev. Biol.* **87**, 286–300 (1981).
- Hedgecock, E. M., Culotti, J. G., Hall, D. H. & Stern, B. D. Genetics of cell and axon migrations in *Caenorhabditis elegans*. *Development* **100**, 365–382 (1987).
- Coulson, A., Sulston, J., Brenner, S. & Karn, J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **83**, 7821–7825 (1986).
- Way, J. C. & Chalfie, M. *mec-3*, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans*. *Cell* **54**, 5–16 (1988).
- Krause, M. & Hirsh, D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* **49**, 753–761 (1987).
- Stringham, E. G., Dixon, D. K., Jones, D. & Candido, E. P. Temporal and spatial expression patterns of the small heat shock (*hsp16*) genes in transgenic *Caenorhabditis elegans*. *Mol. Biol. Cell* **3**, 221–233 (1992).
- Erickson, M., Galletta, B. J. & Abmayr, S. M. *Drosophila myoblast city* encodes a conserved protein that is essential for myoblast fusion, dorsal closure, and cytoskeletal organization. *J. Cell Biol.* **138**, 589–603 (1997).
- Nagase, T. *et al.* Prediction of the coding sequences of unidentified human genes. VI. The coding sequences of 80 new genes (KIAA0201–KIAA0280) deduced by analysis of cDNA clones from cell line KG-1 and brain. *DNA Res.* **3**, 321–329 (1996).
- Mayer, B. J., Hamaguchi, M. & Hanafusa, H. A novel viral oncogene with structural similarity to phospholipase C. *Nature* **332**, 272–275 (1988).
- Clark, E. A. & Brugge, J. S. Integrins and signal transduction pathways: the road taken. *Science* **268**, 233–239 (1995).
- Savill, J., Dransfield, I., Hogg, N. & Haslett, C. Vitronectin receptor-mediated phagocytosis of cells undergoing apoptosis. *Nature* **343**, 170–173 (1990).
- Rushton, E., Drysdale, R., Abmayr, S. M., Michelson, A. M. & Bate, M. Mutations in a novel gene, myoblast city, provide evidence in support of the founder cell hypothesis for *Drosophila* muscle development. *Development* **121**, 1979–1988 (1995).
- Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
- Fire, A., Harrison, S. W. & Dixon, D. A modular set of *lacZ* fusion vectors for studying gene expression in *Caenorhabditis elegans*. *Gene* **93**, 189–198 (1990).
- Robertson, A. G. & Thomson, J. N. Morphology of programmed cell death in the ventral nerve cord of *C. elegans* larvae. *J. Embryo. Exp. Morph.* **67**, 89–100 (1982).
- Yuan, J., Shaham, S., Ledoux, S., Ellis, H. M. & Horvitz, H. R. The *C. elegans* cell death gene *ced-3* encodes a protein similar to mammalian interleukin-1 β -converting enzyme. *Cell* **75**, 641–652 (1993).
- Hengartner, M. O. & Horvitz, H. R. *C. elegans* cell survival gene *ced-9* encodes a functional homolog of the mammalian proto-oncogene *bcl-2*. *Cell* **76**, 665–676 (1994).
- Zou, H., Henzel, W. J., Liu, X., Lutschg, A. & Wang, X. Apaf-1, a human protein homologous to *C. elegans* CED-4, participates in cytochrome c-dependent activation of caspase-3. *Cell* **90**, 405–413 (1997).
- Mello, C. C., Kramer, J. M., Stinchcomb, D. & Ambros, V. Efficient gene transfer in *C. elegans*: Extrachromosomal maintenance and integration of transforming sequences. *EMBO J.* **10**, 3959–3970 (1992).
- Bloom, L. & Horvitz, H. R. The *Caenorhabditis elegans* gene *unc-76* and its human homologs define a new gene family involved in axonal outgrowth and fasciculation. *Proc. Natl Acad. Sci. USA* **94**, 3414–3419 (1997).
- Barstead, R. J. & Waterston, R. H. The basal component of the nematode dense-body is vinculin. *J. Biol. Chem.* **264**, 10177–10185 (1989).
- Okkema, P. G. & Fire, A. The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120**, 2175–2186 (1994).

Acknowledgements. We thank members of H.R.H.'s laboratory and C.-L. Wei for their comments, D. Hall and E. Hedgecock for sharing unpublished results regarding DTC migration, R. Barstead and P. Okkema for cDNA libraries, A. Coulson for cosmid, H. Hasegawa and M. Matsuda for the DOCK180 cDNA clone, E. James for determining DNA sequences of the RACE product, S. Glass for the *ced-5* alleles *n2098* and *n2099* and J. Harris and C. Kenyon for *mu57*. H.R.H. is an Investigator of the Howard Hughes Medical Institute. The accession number for the *ced-5* sequence is AF038576.

Correspondence and requests for materials should be addressed to H.R.H. (e-mail: horvitz@mit.edu).

Human CD14 mediates recognition and phagocytosis of apoptotic cells

Andrew Devitt[†], Odette D. Moffatt[†],
Chandra Raykundalia^{*}, J. Donald Capra[‡],
David L. Simmons[§] & Christopher D. Gregory^{*†}

^{*} Department of Immunology, University of Birmingham Medical School, Birmingham B15 2TT, UK

[†] Institute of Cell Signalling and School of Biomedical Sciences, University of Nottingham Medical School, Queen's Medical Centre, Nottingham NG7 2UH, UK

[‡] Oklahoma Medical Research Foundation, 825 NE 13th Street, Oklahoma City, Oklahoma 73104, USA

[§] SmithKline Beecham Pharmaceuticals, New Frontiers Science Park North, Harlow, Essex CM19 5AW, UK

Cells undergoing programmed cell death (apoptosis) are cleared rapidly *in vivo* by phagocytes without inducing inflammation¹. Here we show that the glycosylphosphatidylinositol-linked plasma-membrane glycoprotein CD14 (refs 2, 3) on the surface of human macrophages is important for the recognition and clearance of apoptotic cells. CD14 can also act as a receptor that

binds bacterial lipopolysaccharide (LPS), triggering inflammatory responses⁴. Overstimulation of CD14 by LPS can cause the often fatal toxic-shock syndrome^{5,6}. Here we show that apoptotic cells interact with CD14, triggering phagocytosis of the apoptotic cells. This interaction depends on a region of CD14 that is identical to, or at least closely associated with, a region known to bind LPS. However, apoptotic cells, unlike LPS, do not provoke the release of pro-inflammatory cytokines from macrophages. These results indicate that clearance of apoptotic cells is mediated by a receptor whose interactions with 'non-self' components (LPS) and 'self' components (apoptotic cells) produce distinct macrophage responses.

We have shown previously that the monoclonal antibody 61D3 binds to the surface of human monocyte-derived macrophages and markedly inhibits their capacity to interact with various leukocytes undergoing apoptosis⁷. We now use transient expression cloning in COS cells⁸ to identify the molecule bearing the epitope that binds to 61D3. The nucleic acid sequence of the complementary DNA identified with this monoclonal antibody (61D3 cDNA) was identical to that of human CD14 (data not shown). Several anti-CD14 mouse monoclonal antibodies and a rabbit polyclonal anti-CD14 antiserum became reactive in cell-binding (not shown) or western blot (Fig. 1) assays when COS cells were transfected with 61D3 cDNA. The protein identified with the 61D3 monoclonal antibody has a relative molecular mass of ~55,000 (*M*_r 55K), the reported size of CD14, and 61D3 also reacts specifically with purified recombinant CD14 (Fig. 1). These results indicate that the 61D3-binding

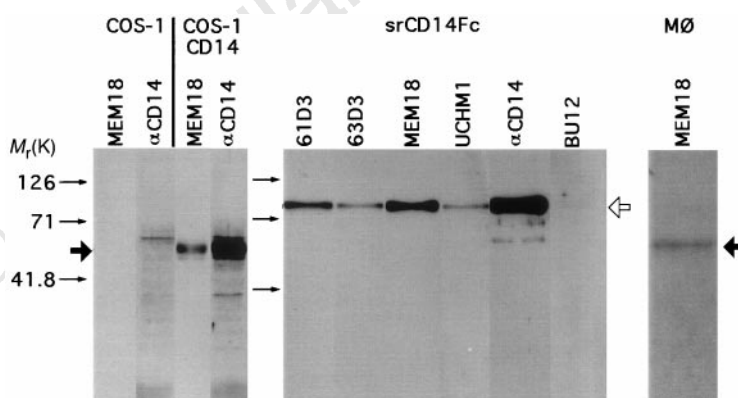


Figure 1 CD14 carries the epitope defined by monoclonal antibody 61D3. Western blot analysis of COS-1 cells transiently transfected with 61D3 cDNA (COS-1/CD14) or control vector (COS-1) (left-hand panel), soluble recombinant CD14-Fc fusion protein (srCD14-Fc) (middle panel), and macrophage (seven-day-

old monocyte-derived) lysate (MØ) (right-hand panel). Blots were probed with the indicated monoclonal antibodies or with a polyclonal rabbit anti-human CD14 antiserum (αCD14). The reported *M*_r of CD14 is 55K (thick black arrows) and that of CD14-Fc is 80K (open arrow).

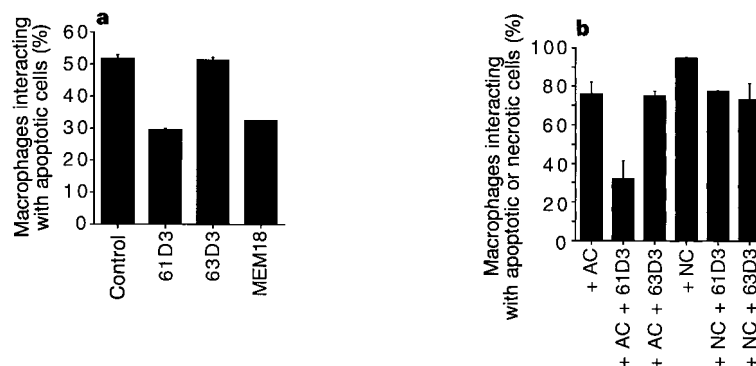


Figure 2 CD14-mediated recognition and phagocytosis of apoptotic cells by macrophages. **a**, Interaction (binding and phagocytosis) of apoptotic Burkitt lymphoma cells with seven-day-old monocyte-derived macrophages, and inhibition of the interaction by monoclonal antibodies 61D3 and MEM18. **b**,

Comparison of the interaction of apoptotic lymphoma cells (AC) and their necrotic counterparts (NC) with macrophages. Inhibition of interaction of AC, but not NC, with macrophages by 61D3.

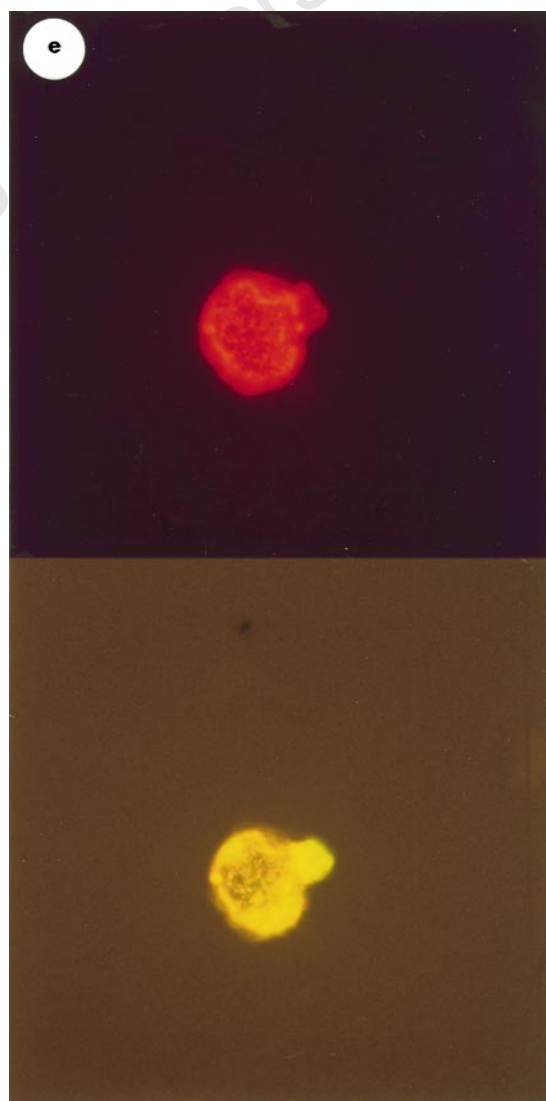
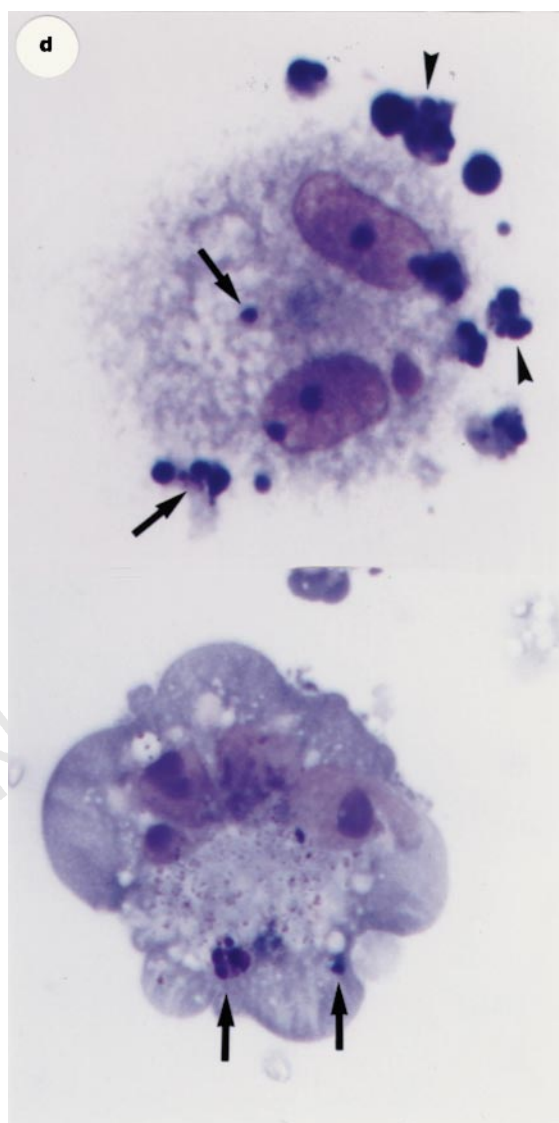
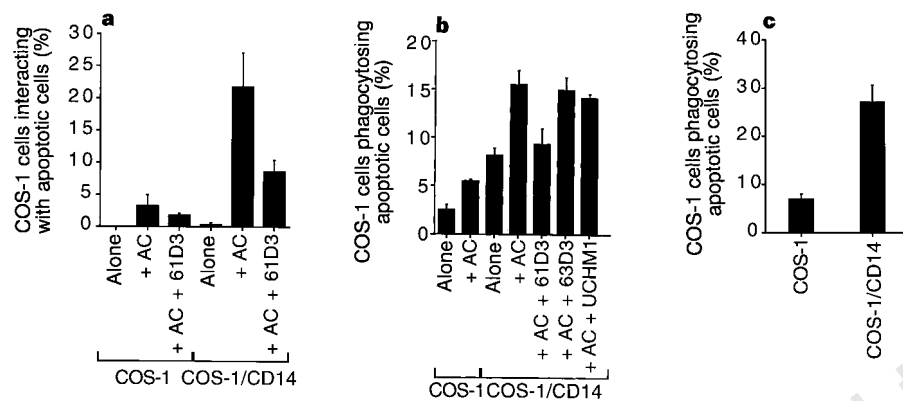


Figure 3 Role of CD14 in recognition and phagocytosis of apoptotic cells. **a**, Increased interaction (binding and phagocytosis) of apoptotic lymphoma cells (AC) with COS-1 cells transiently transfected with CD14 (COS-1/CD14) (31% CD14-positive by immunofluorescence). Inhibition of this interaction by monoclonal antibody 61D3. **b**, Increased phagocytosis of apoptotic cells by COS-1/CD14 (22% CD14-positive) and inhibition of phagocytosis by 61D3. **c**, Fluorescence assay showing enhanced phagocytosis of apoptotic cells by COS-1/CD14

cells. Means $\pm$ s.e.m. of two separate experiments. **c**, COS-1/CD14 cells (Jenner-Giemsa stain). Upper panel, binding and phagocytosis of apoptotic lymphoma cells; lower panel, phagocytosis only. Arrowheads indicate surface-bound apoptotic cells; arrows indicate probable phagocytosed apoptotic cells and debris. **e**, CD14-expressing (red immunofluorescence, upper panel) COS-1 cell engaged in phagocytosis of PKH2-labelled apoptotic lymphoma cells (appearing yellow, lower panel).

epitope is carried by CD14 and that no extra cell-surface molecules, macrophage-associated or otherwise, make a necessary contribution to the epitope.

To understand further how CD14 interacts with apoptotic cells, we investigated the ability of additional anti-CD14 monoclonal antibodies to inhibit the interaction of apoptotic lymphocytes with macrophages (Fig. 2). We also tested the ability of CD14 to support binding and phagocytosis of apoptotic cells when transiently expressed in COS cells (Fig. 3). As shown in Fig. 2a, the anti-CD14 monoclonal antibodies 61D3 and MEM-18, but not 63D3, substantially inhibited binding and phagocytosis of apoptotic lymphocytes by seven-day-old monocyte-derived macrophages. The CD14-mediated interaction of lymphocytes with macrophages was specific for apoptotic cells, as viable lymphocytes did not interact with macrophages in this assay (data not shown; see ref. 7) and interaction of macrophages with necrotic cells was not inhibited by 61D3 (Fig. 2b). When transiently expressed on the surface of COS cells, CD14 was found to impart an increase in both binding (Fig. 3a, d) and phagocytosis (Fig. 3b–e) of apoptotic lymphocytes added to the cultures. CD14 also promoted the phagocytosis of apoptotic COS cells present in the cultures (Fig. 3b). Both binding and phagocytosis of apoptotic lymphocytes by CD14-transfected COS cells were markedly inhibited by the 61D3 monoclonal antibody (Fig. 3a, b). In contrast, phagocytosis by CD14 transfectants was unaffected by the anti-CD14 monoclonal antibodies 63D3 and UCHM-1.

A two-colour fluorescence assay using anti-CD14 monoclonal antibody and fluorescent apoptotic cells showed that enhanced phagocytosis occurred specifically in CD14-expressing COS cells (Fig. 3c, e). These results indicate that CD14 supports both binding and phagocytosis of apoptotic cells, and that this activity is not dependent on a macrophage context. As only 25–30% of CD14-expressing COS cells were observed to phagocytose apoptotic cells (despite overexpression of the molecule on COS cells as compared with macrophages; data not shown), we suggest that additional molecules contribute to this activity in the context of the professional phagocyte.

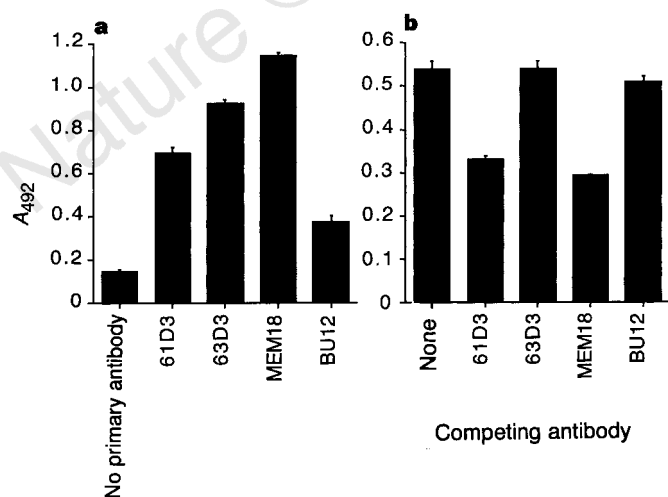


Figure 4 Epitope mapping of the 61D3 monoclonal antibody by ELISA, using immobilized soluble recombinant CD14-Fc fusion protein. **a**, CD14-binding of murine anti-CD14 monoclonal antibodies 61D3 (10 $\mu\text{g ml}^{-1}$), 63D3 (10 $\mu\text{g ml}^{-1}$), MEM-18 (1:50) and BU12 (CD19, 10 $\mu\text{g ml}^{-1}$), visualized using anti-mouse-Ig-HRP. **b**, Competition for binding to CD14 of biotinylated 61D3 (5 ng ml^{-1}) by the indicated monoclonal antibodies. Binding was visualized using streptavidin-HRP. In the experiment shown, binding of 61D3 to CD14 was reduced by 39% and 46% in the presence of 61D3 and MEM-18, respectively.

The observation that both 61D3 and MEM-18 inhibited the interaction of apoptotic cells with macrophages indicated that these antibodies may recognize neighbouring or identical regions of CD14. To address this question, we used a competitive enzyme-linked immunosorbent assay (ELISA) of the binding of monoclonal antibodies to recombinant CD14 (Fig. 4). Monoclonal antibodies 61D3, 63D3 and MEM-18 were all strongly reactive with recombinant CD14 in the ELISA (Fig. 4a). However, whereas 61D3 and MEM-18 strongly inhibited the binding of further 61D3 to the recombinant molecule, 63D3, an antibody that lacked blocking activity in the macrophage assay, did not (Fig. 4b). Conversely, 63D3, but not 61D3 or MEM-18, strongly inhibited the binding of further 63D3 to recombinant CD14 (data not shown). Therefore, the activity of CD14 in regulating the interaction of apoptotic cells with macrophages depends on a molecular site that is identical to, or maps close to, the epitope(s) defined by 61D3 and MEM-18.

MEM-18 is known to inhibit potently the binding of CD14 to its prototypic ligand, LPS⁸. Thus the binding sites on CD14 for LPS and for apoptotic cells are either identical or closely associated. Apoptotic cells are thought to be cleared by macrophages without the induction of pro-inflammatory responses. However, LPS clearance through CD14 activates macrophages to release pro-inflammatory mediators. We therefore determined whether pro-inflammatory cytokine release was triggered in macrophage populations that interact with apoptotic cells through CD14. Figure 5 shows the production of tumour-necrosis factor- α (TNF- α) by seven-day-old monocyte-derived macrophages, which use CD14 to interact with apoptotic cells (Fig. 2a). These macrophage cultures did not produce TNF- α in response to apoptotic cells. In contrast, exposure of the same seven-day-old macrophage cultures to LPS resulted in significant TNF- α release (Fig. 5). TNF- α production, like uptake of apoptotic cells, was inhibited by 61D3 and MEM-18 but not by 63D3 (Fig. 5), indicating that apoptotic cells and LPS interact with closely associated or identical regions of CD14. Nevertheless, LPS promotes pro-inflammatory cytokine release whereas apoptotic cells do not.

These findings establish a previously unrecognized function for

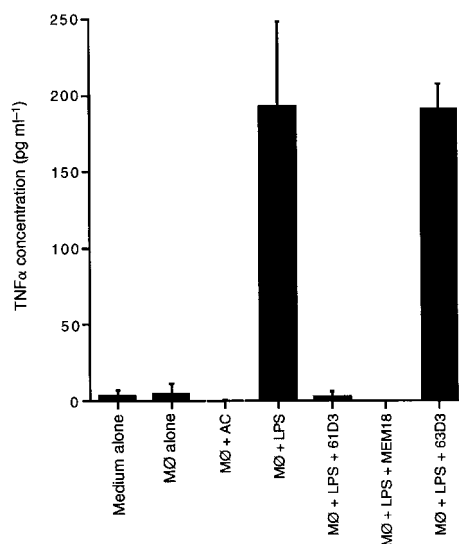


Figure 5 Production of TNF- α by macrophages stimulated with LPS but not apoptotic cells. Macrophages (MØ) were incubated either alone or in the presence of apoptotic lymphoma cells (AC) or LPS (5 ng ml^{-1}) or LPS with monoclonal antibodies 61D3, MEM-18 or 63D3, for 3 h before assaying the concentration of TNF- α present in the culture supernatants. Background levels of TNF- α present in the assay medium (medium alone) are shown for comparison.

CD14 and extend to five, with the vitronectin receptor $\alpha_v\beta_3$ integrin⁹, CD36 (ref. 10), class A scavenger receptor¹¹ and the ATP-binding-cassette transporter ABC-1 (ref. 12), the number of known macrophage surface molecules involved in the removal of apoptotic cells. The apparent abundance of phagocyte receptors for apoptotic cells may illustrate the vital importance of the clearance process. In the absence of clearance, or when it is overwhelmed¹³, secondary necrosis occurs, with the catastrophic consequences of inflammation and damage to normal tissue.

Although interactions between apoptotic cells and phagocytes are most simply envisaged as receptor–ligand pairing at the surface of apposed cells, intermediate molecules such as thrombospondin, which binds $\alpha_v\beta_3$ integrin and CD36 through separate epitopes and appears to form a molecular ‘bridge’ between macrophages and apoptotic cells¹⁰, may also be involved. CD14 interacts with several constituents of microorganisms, including, in addition to LPS, muramyl dipeptide and soluble peptidoglycan¹⁴, uronic acid polymers¹⁵, streptococcal cell-wall polysaccharides¹⁶, mycobacterial lipoarabinomannan¹⁷, and a yeast cell-wall protein, WI-1 (ref. 18). Cell-wall components of both Gram-positive and Gram-negative bacteria^{17,19} bind to CD14. Potentially, therefore, several categories of molecule may become available as CD14 ligands on the surface of apoptotic cells. None of the known cell-surface components of apoptotic cells has been linked specifically with any of the known macrophage receptors for apoptotic cells, although carbohydrate^{20,21} and phospholipid²² changes at the cell surface during apoptosis may yield important ligands. The anionic phospholipid phosphatidylserine is redistributed to the outer leaflet of the plasma membrane during apoptosis and is involved in the recognition of apoptotic cells by certain classes of macrophage²². Phosphatidylserine may provide an apoptotic-cell-associated ligand for CD14, as suggested by a study in which the 61D3 monoclonal antibody blocked the (phosphatidylserine-dependent) interaction of lipid-symmetric erythrocytes with macrophages²³. CD14 interacts with anionic phospholipids²⁴ and acts as a lipid-transfer protein that may play a role in lipid-A-dependent LPS responses²⁵. In addition, the ability of CD14 to bind polysaccharide-containing molecules²⁶ indicates that its interaction with apoptotic cells may occur through a lectin-like activity. This would concur with the proposal²⁰ that the recognition of apoptotic cells by phagocytes involves interactions between sugars on the altered apoptotic-cell surface and novel macrophage-surface-associated lectins.

We conclude that CD14 is a multifunctional receptor which, in addition to its known role as an innate immune receptor for ‘infectious non-self’ components¹⁷, also interacts with ‘apoptotic self’ components. Further work is required to find the important CD14 ligand(s) on apoptotic cells and to define the mechanism underlying the different macrophage responses—non-inflammatory apoptotic-cell clearance versus pro-inflammatory cytokine release—to CD14 ligation. It will also be interesting to determine whether apoptotic cell/CD14 interactions generate signals that actively suppress pro-inflammatory signals, as suggested by recent work^{27,28}. The differential signalling following ligation of CD14 by LPS and apoptotic cells is consistent with the proposal that ligand-induced association of this glycosylphosphatidylinositol-linked receptor with transmembrane proteins^{4,29} is a likely mechanism of initiation of intracellular signal-transduction cascades. Our finding that CD14 does not appear to function as efficiently in the COS cell as it does in the macrophage (Figs 2 and 3) also indicates that additional macrophage molecules may cooperate with CD14 to effect the binding and phagocytosis of apoptotic cells. □

Methods

Expression cloning. Transient expression cloning was carried out as described³ using a cDNA library in plasmid pCDM8, from the human promyelocytic leukaemia cell line HL60 stimulated with dibutyryl cAMP. After three rounds of panning using the 61D3 monoclonal antibody³⁰, plasmid

DNA was purified from 12 individual clones and transiently expressed in COS-1 cells. Reactivity with 61D3 monoclonal antibody was detected in one clone by indirect immunofluorescence, and complete sequencing of this clone (61D3 cDNA) was carried out by the Biopolymer Synthesis and Analysis Unit, University of Nottingham Medical School, using a Perkin Elmer/ABI 373A sequencer. The 61D3 cDNA sequence was identical to that of human CD14 cDNA, as determined using the GCG package to search the GenBank and EMBL nucleotide databases (GCG: <http://ictsg10.unil.ch:8080/w2h/>; EMBL: gopher://biox.emblnet.unibas.ch:13021/77/index/embl/index; GenBank: <http://www.ncbi.nlm.nih.gov/web/genbank/index.html>).

Immunofluorescence staining, flow cytometry and immunoblotting. The following monoclonal antibodies were used: 61D3, 63D3 (ref. 30), MEM-18 (Monosan), UCHM-1 (Sigma), TÜK4 (Dako) (all anti-CD14) and BU12 (anti-CD19). Indirect immunofluorescence staining was carried out on COS-1 cells in suspension and flow-cytometric analysis was performed using a FACScan (Becton Dickinson). For immunoblotting, cell lysates or protein preparations were separated using one-dimensional 10% SDS–PAGE. Proteins were blotted to PVDF membrane (Immobilon P, Millipore) and detected using the ECL system (Amersham).

Production of soluble recombinant CD14-Fc and anti-CD14 antiserum.

CD14 cDNA³ subcloned into the plg vector (Invitrogen) was transfected into COS-1 cells. Expression and secretion of the CD14-Fc fusion protein was allowed to proceed for 5 days before affinity purification, using a protein A Sepharose column (Pharmacia Biotech), from the culture supernatant. Polyclonal anti-CD14 antibody was raised in rabbits immunized with CD14-Fc fusion protein. Anti-Fc reactivity was removed by absorption with human IgG.

Macrophage and COS-1 recognition/phagocytosis assays. Macrophages were obtained from peripheral blood monocytes of normal healthy donors. Seven-day-old monocyte-derived macrophages cultured on multiwell glass slides (Hendley, Essex, UK) were incubated with apoptotic cells for 1 h in the presence or absence of monoclonal antibodies as described⁷. Macrophages were washed extensively to remove unbound apoptotic cells, fixed in methanol and stained with Jenner-Giemsa. Cultures were scored by light microscopy for the percentage of macrophages interacting with apoptotic cells, both surface-bound and phagocytosed, using established criteria⁷. Interaction of apoptotic leukocytes with macrophages from donors used in these experiments was routinely inhibited by 61D3 monoclonal antibody (>90% of donors).

COS-1 cells were transfected as described³ with 61D3 cDNA in plasmid pCDM8 and cultured on multiwell glass slides for 3 days. Mock-transfected COS-1 cells were used as controls. Co-incubation with apoptotic cells, removal of unbound cells and Jenner-Giemsa staining were carried out as for the macrophage recognition/phagocytosis assay. COS cells interacting (that is, binding and/or phagocytosing) with two or more apoptotic cells were scored as positive.

To assess phagocytosis of apoptotic cells separately from cell–cell binding, COS-1 cells with associated apoptotic cells were treated with 0.05% trypsin/0.02% EDTA for 2 min. Cyto-centrifuge smears of the resultant cell suspensions were made and stained with Jenner-Giemsa. In these preparations, no apoptotic cells bound to the surface of COS-1 cells were apparent. Simultaneous assessment of phagocytic activity and CD14 expression by COS-1 cells was achieved using a fluorescence assay. Before the induction of apoptosis, lymphocytes were labelled using the PKH2 General Cell Linker Kit (Sigma). Labelling was carried out using 5 μ M PKH2 for 5 min. COS-1 cells transfected with 61D3 cDNA were exposed to PKH2 (green)-labelled apoptotic lymphocytes, subjected to trypsin/EDTA treatment as described above, and stained by immunofluorescence for surface CD14 (red) using 63D3 and phycoerythrin-conjugated goat anti-mouse IgG (Sigma). Assays were scored by fluorescence microscopy of 1,000 CD14-positive and 1,000 CD14-negative cells from individual transfections.

In recognition/phagocytosis assays, the group I Burkitt lymphoma cell line Mutu I, induced into apoptosis by 16 h incubation with 1 μ g ml^{−1} of the calcium ionophore ionomycin (Calbiochem)⁷, was used as the source of apoptotic cells. Viable cells from this line fail to interact with macrophages⁷ or COS-1 cells (data not shown) in these assays. Necrosis was induced by heat treatment of Mutu I cells at 56 °C for 40–45 min until they were 60–70% trypan-blue-positive. All binding and phagocytosis assays were carried out in the presence of human serum. Where indicated, monoclonal antibodies were

included at the following concentrations: 61D3 and 63D3, 80 $\mu\text{g ml}^{-1}$; MEM-18 and UCHM-1, 1:20 dilution of manufacturer's preparations.

CD14 epitope mapping. Binding of monoclonal antibodies to CD14 was assessed by ELISA using CD14-Fc (5 $\mu\text{g ml}^{-1}$) captured with immobilized sheep anti-human IgG (The Binding Site, 10 $\mu\text{g ml}^{-1}$). Bound antibody was visualized using sheep anti-mouse immunoglobulin (The Binding Site) absorbed against normal human serum and conjugated with horseradish peroxidase (HRP). In competition assays, binding of biotinylated 61D3 or biotinylated 63D3 to captured CD14-Fc in the presence or absence of competing monoclonal antibodies was detected using streptavidin-HRP (Southern Biotechnology), *o*-phenylenediamine dihydrochloride substrate and an A_{492} read-out.

TNF- α assay. TNF- α was measured in macrophage-culture supernatants after treatment of macrophages for 3 h with either LPS (*Escherichia coli* O111:B4, Sigma) or apoptotic cells in the presence or absence of monoclonal antibodies. TNF- α was assayed by ELISA using matched-pair capture and biotinylated detection antibodies (R&D Systems).

Unless otherwise stated, all data shown are means $\pm$ s.e.m., with experiments representative of at least three similar experiments.

Received 3 February; accepted 2 March 1998.

- Savill, J., Fadok, V., Henson, P. & Haslett, C. Phagocyte recognition of cells undergoing apoptosis. *Immunol. Today* **14**, 131–136 (1993).
- Goyert, S. M. *et al.* The CD14 monocyte differentiation antigen maps to a region encoding growth factors and receptors. *Science* **239**, 497–500 (1988).
- Simmons, D. L., Tan, S., Tenen, D. G., Nicholson-Weller, A. & Seed, B. Monocyte antigen CD14 is a phospholipid anchored membrane protein. *Blood* **73**, 284–289 (1989).
- Ulevitch, R. J. & Tobias, P. S. Receptor-dependent mechanisms of cell stimulation by bacterial endotoxin. *Annu. Rev. Immunol.* **13**, 437–457 (1995).
- Ferrero, E. *et al.* Transgenic mice expressing human CD14 are hypersensitive to lipopolysaccharide. *Proc. Natl Acad. Sci. USA* **90**, 2380–2384 (1993).
- Haziot, A. *et al.* Resistance to endotoxin-shock and reduced dissemination of Gram-negative bacteria in CD14-deficient mice. *Immunity* **4**, 407–414 (1996).
- Flora, P. K. & Gregory, C. D. Recognition of apoptotic cells by human macrophages-inhibition by a monocyte/macrophage-specific monoclonal antibody. *Eur. J. Immunol.* **24**, 2625–2632 (1994).
- Juan, T. S. C. *et al.* Identification of a lipopolysaccharide binding domain in CD14 between amino acids 57 and 64. *J. Biol. Chem.* **270**, 5219–5224 (1995).
- Savill, J., Dransfield, I., Hogg, N. & Haslett, C. Vitronectin receptor-mediated phagocytosis of cells undergoing apoptosis. *Nature* **343**, 170–173 (1990).
- Savill, J., Hogg, N., Ren, Y. & Haslett, C. Thrombospondin cooperates with CD36 and the vitronectin receptor in macrophage recognition of neutrophils undergoing apoptosis. *J. Clin. Invest.* **90**, 1513–1522 (1992).
- Platt, N., Suzuki, H., Kurihara, Y., Kodama, T. & Gordon, S. Role for the class A macrophage scavenger receptor in the phagocytosis of apoptotic thymocytes *in vitro*. *Proc. Natl Acad. Sci. USA* **93**, 12456–12460 (1996).
- Luciani, M.-F. & Chimini, G. The ATP binding cassette transporter ABC1, is required for the engulfment of corpses generated by apoptotic cell death. *EMBO J.* **15**, 226–235 (1996).
- Ogasawara, J. *et al.* Lethal effect of the anti-Fas antibody in mice. *Nature* **364**, 806–809 (1993).
- Weidemann, B. *et al.* Specific binding of soluble peptidoglycan and muramyl dipeptide to CD14 on human monocytes. *Infect. Immun.* **65**, 858–864 (1997).
- Espevik, T. *et al.* The involvement of CD14 in stimulation of cytokine production by uronic acid polymers. *Eur. J. Immunol.* **23**, 255–261 (1993).
- Soell, M. *et al.* Activation of human monocytes by streptococcal rhamnose glucose polymers is mediated by CD14 antigen, and mannan-binding protein inhibits TNF- α release. *J. Immunol.* **154**, 851–860 (1995).
- Pugin, J. *et al.* CD14 is a pattern recognition receptor. *Immunity* **1**, 509–516 (1994).
- Newman, S. L., Chaturvedi, S. & Klein, B. S. The WI-1 antigen of *Blastomyces dermatitidis* yeasts mediates binding to human macrophage CD11b/CD18 (CR3) and CD14. *J. Immunol.* **154**, 753–761 (1995).
- Wright, S. D., Ramos, R. A., Tobias, P. S., Ulevitch, R. J. & Mathison, J. C. CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein. *Science* **249**, 1431–1433 (1990).
- Duvall, E., Wyllie, A. H. & Morris, R. G. Macrophage recognition of cells undergoing programmed cell death (apoptosis). *Immunology* **56**, 351–358 (1985).
- Dini, L., Autuori, F., Lentini, A., Oliverio, S. & Piacentini, M. The clearance of apoptotic cells in the liver is mediated by the asialoglycoprotein receptor. *FEBS Lett.* **296**, 174–178 (1992).
- Fadok, V. A. *et al.* Exposure of phosphatidylserine on the surface of apoptotic lymphocytes triggers specific recognition and removal by macrophages. *J. Immunol.* **148**, 2207–2216 (1992).
- Pradhan, D., Krahling, S., Williamson, P. & Schlegel, R. A. Multiple systems for recognition of apoptotic lymphocytes by macrophages. *Mol. Biol. Cell* **8**, 767–778 (1997).
- Yu, B., Hailman, E. & Wright, S. D. Lipopolysaccharide binding protein and soluble CD14 catalyze exchange of phospholipids. *J. Clin. Invest.* **99**, 315–324 (1997).
- Delude, R. L. *et al.* CD14 enhances cellular responses to endotoxin without imparting ligand-specific recognition. *Proc. Natl Acad. Sci. USA* **92**, 9288–9292 (1995).
- Cavallion, J. M. *et al.* CD14/LPS receptor exhibits lectin-like properties. *J. Endotox. Res.* **3**, 471–480 (1996).
- Voll, R. E. *et al.* Immunosuppressive effects of apoptotic cells. *Nature* **390**, 350–351 (1997).
- Fadok, V. A. *et al.* Macrophages that have ingested apoptotic cells *in vitro* inhibit proinflammatory cytokine production through autocrine/paracrine mechanisms involving TGF- β , PGE2 and PAF. *J. Clin. Invest.* **101**, 890–898 (1998).
- Petty, H. R. & Todd, R. F. Integrins as promiscuous signal transduction devices. *Immunol. Today* **17**, 209–212 (1996).
- Ugolini, V., Nunez, G., Smith, R. G., Stastny, P. & Capra, J. D. Initial characterization of monoclonal antibodies against human monocytes. *Proc. Natl Acad. Sci. USA* **77**, 6764–6768 (1980).

Acknowledgements. This work was supported by the MRC.

Correspondence and requests for materials should be addressed to C.D.G. (e-mail: Chris.Gregory@nottingham.ac.uk).

Inhibition of oxytocin receptor function by direct binding of progesterone

Eric Grazzini*, Gilles Guillon†, Bernard Mouillac† & Hans H. Zingg*

* Laboratory of Molecular Endocrinology, Royal Victoria Hospital Research Institute, McGill University, Montreal, Québec H3A 1A1, Canada

† Unité INSERM U469, Centre de Pharmacologie-Endocrinologie, Montpellier 34094, France

The steroid hormone progesterone (P_4) is essential for establishing and maintaining pregnancy in mammals^{1–3}. One of its functions includes maintenance of uterine quiescence by decreasing uterine sensitivity to the uterotonic peptide hormone oxytocin^{3–5}. Although it is generally held that steroid hormones such as P_4 act at a genomic level by binding to nuclear receptors and modulating the expression of specific target genes⁶, we show here that the effect of P_4 on uterine sensitivity to oxytocin involves direct, non-genomic action of P_4 on the uterine oxytocin receptor (OTR), a member of the G-protein-coupled receptor family. P_4 inhibits oxytocin binding to OTR-containing membranes *in vitro*, binds with high affinity to recombinant rat OTR expressed in CHO cells, and suppresses oxytocin-induced inositol phosphate production and calcium mobilization. These effects are highly steroid- and receptor-specific, because binding and signalling functions of the closely related human OTR are not affected by P_4 itself but by the P_4 metabolite 5 β -dihydroprogesterone. Our findings provide the first evidence for a direct interaction between a steroid hormone and a G-protein-coupled receptor and define a new level of crosstalk between the peptide- and steroid-hormone signalling pathways.

Because P_4 -induced reduction in uterine oxytocin binding occurs in the absence of protein synthesis⁵ and is not accompanied by a significant decrease in OTR gene expression⁷, we investigated whether P_4 could affect OTR function through a non-genomic mechanism. As shown in Fig. 1a, addition of P_4 *in vitro* to membranes derived from a parturient rat uterus inhibited binding of the OTR-specific ligand OTA⁸ (inhibition constant, $K_i = 16 \pm 2$ nM; maximal inhibition was $59 \pm 6\%$ at 1 μM). As circulating P_4 concentrations in the rat reach 500 nM during pregnancy⁹, our effective P_4 concentrations were within physiological range. Inhibition was due to a decrease in binding capacity ($B_{\text{max}} = 0.80 \pm 0.02$, compared with 0.39 ± 0.01 pmol mg^{-1} protein) without an effect on binding affinity ($K_D = 0.23 \pm 0.07$ compared with 0.20 ± 0.03 nM) (Fig. 1b).

P_4 also induced a dose-dependent reduction of specific oxytocin binding to recombinant rat OTR expressed in CHO cells ($K_i = 15 \pm 2$ nM; Fig. 1c and d). By contrast, P_4 had no effect on binding of the V_{1a} vasopressin-receptor-specific ligand LVA¹⁰ to recombinant V_{1a} receptor ($V_{1a}R$) (Fig. 1c). Similar results were obtained with ³H-labelled natural agonists ($K_i = 19 \pm 3$ nM for [³H]oxytocin; no inhibition of [³H]vasopressin binding). P_4 coupled to bovine serum albumin also reduced oxytocin binding

Table 1 Effects of P_4 on [³H]AVP binding to OTR and $V_{1a}R$

Receptor	P_4	B_{max} (pmol mg^{-1} protein)	Inhibition by P_4 (%)
OTR	–	0.96 ± 0.08	
	+	0.37 ± 0.03	61 ± 6
$V_{1a}R$	–	1.32 ± 0.05	
	+	1.31 ± 0.04	1 ± 3

Results are the means $\pm$ s.e. of triplicate determinations.

to intact OTR-expressing CHO cells ($K_i = 85 \pm 8$ nM), thereby excluding a cytoplasmic site of P_4 action.

To determine whether the P_4 effect was ligand- or receptor-specific, we took advantage of the fact that arginine vasopressin (AVP) binds to both the OTR and the $V_{1a}R$ with comparable affinity¹¹. [3H]AVP binding to the OTR was repressed by P_4 , whereas [3H]AVP binding to $V_{1a}R$ was unaffected (Table 1), indicating that the specificity of the P_4 effect is determined by the receptor and not the ligand.

P_4 bound with high affinity to recombinant OTR expressed in CHO cells ($K_d = 20 \pm 3$ nM, $B_{max} = 2.7 \pm 0.9$ pmol mg^{-1} protein; Fig. 2a). No P_4 binding sites were present in wild-type CHO cells or in CHO cells expressing the $V_{1a}R$. P_4 and oxytocin binding to the OTR were reciprocally regulated by the state of G-protein coupling to the receptor. G-protein uncoupling induced by the non-hydrolysable analogue GTP γ S had no effect on the binding of competitive antagonist (^{125}I -labelled OTA) but upregulated specific [3H] P_4 binding and reduced agonist ([3H]oxytocin) binding (Fig. 2b).

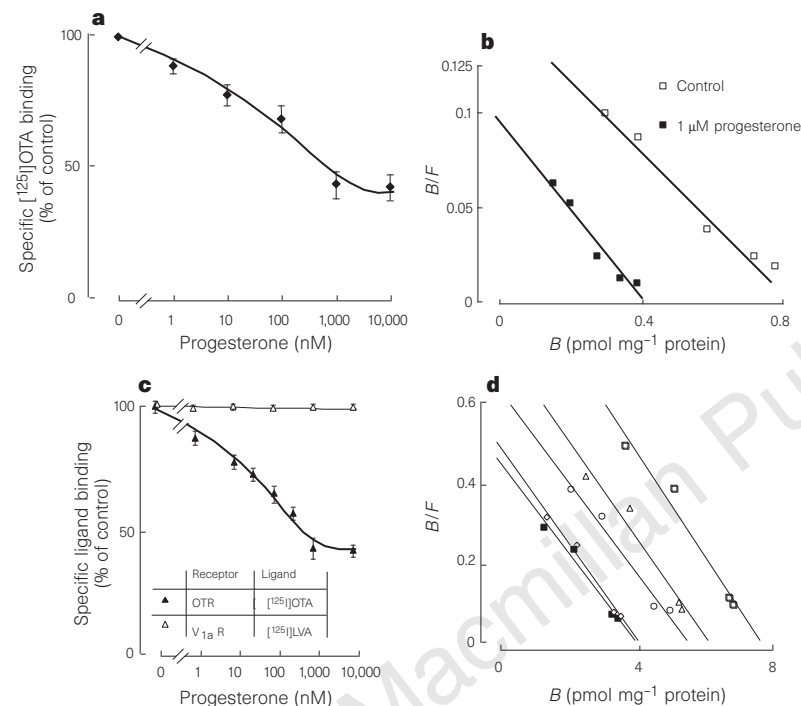


Figure 1 Inhibition of OTR binding capacity by P_4 . **a**, Inhibition by P_4 of specific binding of [^{125}I]OTA (0.2 nM) to parturient rat uterine membranes. **b**, Scatchard plot of [^{125}I]OTA binding to parturient rat uterine membranes in the absence or presence of P_4 . B, amount bound; F, amount free. **c**, Inhibition by P_4 of [^{125}I]OTA or [^{125}I]LVA binding to recombinant OTR or $V_{1a}R$, respectively, expressed in CHO cells. **d**, Scatchard plots of [^{125}I]OTA binding to membranes of OTR-expressing CHO cells in the presence of vehicle ($\square$), 0.01 μ M (Δ), 0.1 μ M ($\circ$), 1.0 μ M ($\diamond$) or 10 μ M ($\blacksquare$) P_4 .

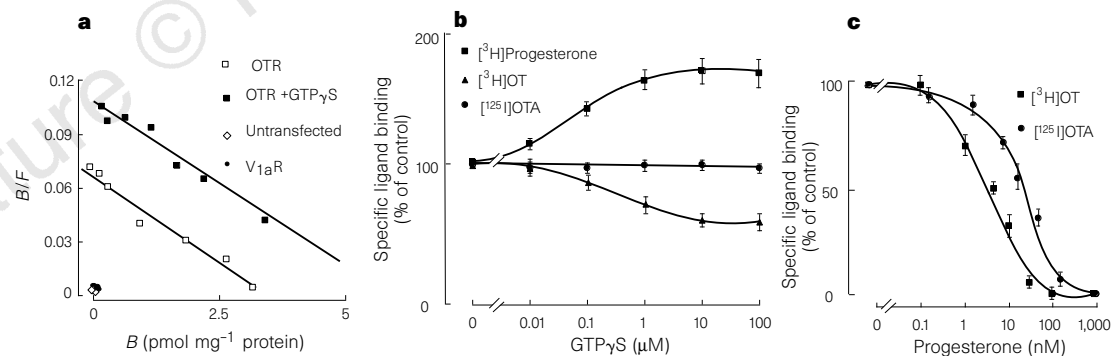


Figure 2 Specific binding of P_4 to the OTR and effect of GTP γ S. **a**, Scatchard plot of [3H] P_4 binding to membranes derived from CHO cells expressing the OTR in the presence or absence of 10 μ M GTP γ S. No binding was observed to untransfected cells or to cells expressing the $V_{1a}R$. **b**, GTP γ S effects on agonist [3H]OT, antagonist [^{125}I]OTA, 0.2 nM or [3H] P_4 (20 nM) binding to the OTR. **c**, P_4 -induced inhibition of agonist [3H]OT, 2 nM or antagonist [^{125}I]OTA, 0.2 nM binding to the OTR in presence of 10 μ M GTP γ S.

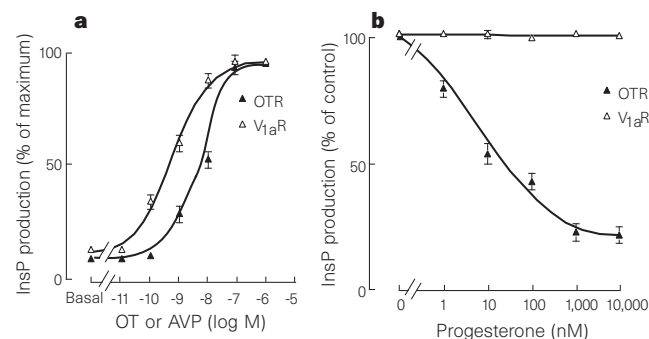


Figure 3 Effect of P_4 on ligand-induced IP production. **a**, Effect of OT ($\blacktriangle$) or AVP ($\triangle$) on InsP production by CHO cells expressing the OTR or the V_{1a} receptor, respectively (ED₅₀: 2 nM and 0.65 nM, respectively). **b**, Effect of P_4 on OT- ($\blacktriangle$, 2 nM) or AVP- ($\triangle$, 0.65 nM) induced InsP production.

The upregulation of P_4 binding was due to an increased value of $B_{\max}$ (Fig. 2a). A similar GTP γ S-induced modulation of binding has been observed for inverse agonists^{12–15}. Comparison of Figs 2c and 1c reveals that, consistent with its stimulating effect on P_4 binding, GTP γ S also enhanced the inhibition by P_4 of 3 H-oxytocin and 125 I-labelled OTA binding (100% inhibition at 100 nM and 1 μ M of P_4 for 3 H]oxytocin and 125 I-labelled OTA, respectively). These results indicate the P_4 interactions with the OTR are affected by guanine-nucleotide-induced changes of receptor conformation.

We next investigated the effect of P_4 on receptor-mediated signalling. Oxytocin and AVP stimulated inositol phosphate (InsP) production in a dose-dependent fashion in OTR- and V_{1a} R-expressing CHO cells, respectively (half-maximal effective dose, ED_{50} : 2.0 ± 0.5 nM and 0.65 ± 0.08 nM; Fig. 3a). We investigated the effect of P_4 on receptor signalling in the presence of ligand concentrations corresponding to the ED_{50} . A dose-dependent inhibition of oxytocin-induced InsP accumulation was observed ($K_i = 11 \pm 3$ nM; maximal inhibition: $85 \pm 6\%$ at 1 μ M; Fig. 3b). P_4 did not affect AVP-induced InsP production in V_{1a} R-expressing cells, indicating that the effect on OTR signalling was receptor-specific.

Table 2 Steroid effects on rat OTR binding and signalling

Steroid	[125 I]OTA binding* (% inhibition)	InsP accumulation† (% inhibition)
Progesterone	50.7 ± 7.1	72.3 ± 5.3
Aldosterone	1.8 ± 7.2	7.2 ± 7.4
Testosterone	5.4 ± 5.3	3.6 ± 5.6
Oestrogen	0.7 ± 3.8	3.8 ± 5.6
Dexamethasone	5.3 ± 5.3	3.5 ± 7.2

* Binding to membranes of rat OTR-expressing CHO cells.

† InsP production by rat OTR-expressing CHO cells in response to 2 nM oxytocin.

Table 3 Inhibition of [3 H]oxytocin binding to OTR by P_4 and P_4 derivatives

Steroid compound	K_i (nM)*	
	Rat OTR	Human OTR
P_4	19 ± 3	None†
R5020	59 ± 7	None†
RU486	46 ± 5	None†
5 β -pregnane 3,20-dione (5 β -dihydroprogesterone)	None†	32 ± 5

* Membranes derived from CHO cells expressing either the rat or the human OTR were incubated with 2 nM [3 H]oxytocin and the effect of steroids on oxytocin binding was expressed as the inhibition constant, K_i .

† No inhibition was observable at concentrations up to 10 μ M.

The effect of P_4 on oxytocin-induced changes in intracellular calcium concentration ($[Ca_i]$) was assessed by spectrofluorometric analysis¹⁶. Application of 2 nM oxytocin led to a rapid rise in $[Ca_i]$ in OTR-expressing CHO cells (Fig. 4a). This effect was dose-dependent ($ED_{50} = 2.2 \pm 0.1$ nM; Fig. 4c). Application of 1 μ M P_4 caused a sharp but reversible reduction in the oxytocin-induced calcium response (Fig. 4b, d). The speed of this P_4 effect further supports the idea that P_4 acts by a non-genomic mechanism.

Of the natural steroids tested, only P_4 had an effect on oxytocin binding or InsP production (Table 2); of the synthetic P_4 derivatives tested, progestin R5020 and the antiprogestin RU486 had similar effects (Table 3). Oxytocin binding to the human OTR was unaffected by P_4 but was altered by the P_4 metabolite 5 β -dihydroprogesterone (Table 3). Consistent with this, 5 β -dihydroprogesterone, but not P_4 , inhibited oxytocin-induced InsP production in human OTR-transfected CHO cells ($K_i = 0.25$ nM). These results indicate that the observed effects are highly receptor- and steroid-specific and are unlikely to be due to nonspecific membrane interactions or to interaction with another unknown binding protein.

Non-genomic effects of P_4 or P_4 metabolites include the induction of oocyte maturation¹⁷ and interactions with the GABA $_A$ receptor¹⁸, the NMDA receptor¹⁹, the nicotinic acetylcholine receptor²⁰ and a sperm-cell-membrane P_4 receptor²¹. Our results provide the first demonstration of a functional interaction between a steroid and a member of the large class of G-protein-linked receptors. The data define not only a new type of non-genomic mechanism of steroid action but also a previously undiscovered level of cross-talk between steroid- and peptide-mediated signalling pathways. The fact that the human and the rat OTRs interact with different steroid molecules and the observation that the steroid/receptor interactions are modulated by GTP γ S indicate that there is a direct steroid/OTR interaction. There may be a mechanistic relation between a previously described effect of P_4 on OTR displacement along dendrites²² and the P_4 /OTR interaction described here. The steroid-induced modulation of G-protein-linked receptor signalling shown here may be the first example of other similar interactions and should help in the design of non-peptide receptor antagonists. □

Methods

Tissues and cells. Uteri were removed from timed-pregnant Sprague-Dawley rats (Charles River, Canada) at the moment of parturition. Untransfected CHO

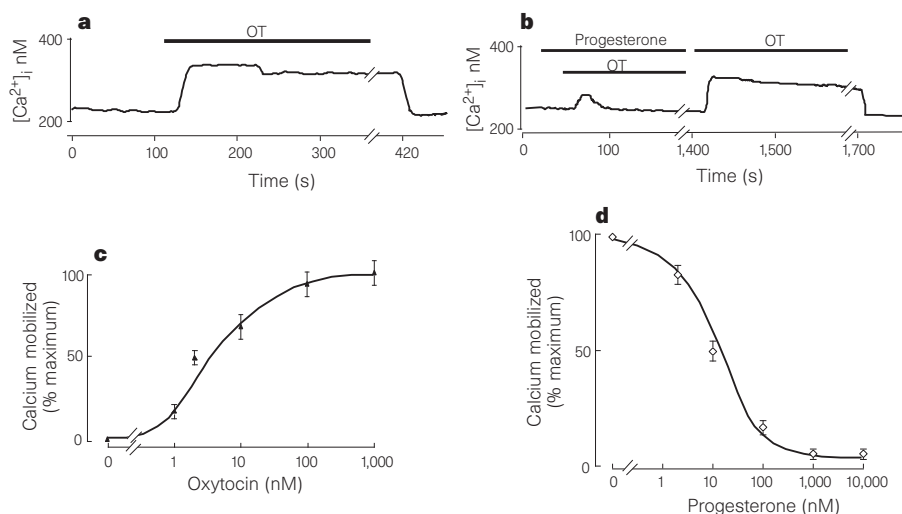


Figure 4 Effect of P_4 on OT-stimulated calcium mobilization in OTR-expressing Fura-2-loaded CHO cells. **a**, $[Ca_i]$ was determined by dual-wavelength spectroscopy^{16,25}. OT (2 nM) was present in the medium during the time indicated by the horizontal bar. **b**, As **a**, but P_4 (1 μ M) was also present in the medium during

the time indicated by the horizontal bar. During the time indicated by the first x-axis break, cells were incubated in medium devoid of P_4 or OT. **c**, Dose-response curve of OT effect on Ca_i mobilization. **d**, Dose-response curve of P_4 effect on 2 nM OT-induced Ca_i mobilization.

cells, CHO cells stably transfected with an expression vector encoding the rat OTR²⁷ (from S. Lolait), the human OTR²³ or the rat V_{1a} receptor²⁴ were grown in α -modified DMEM/5% fetal calf serum.

Binding experiments. Iodination of ligands and binding experiments were done as described¹⁶. OTA⁸ and LVA¹⁰ were obtained from M. Manning. [³H]-oxytocin ([³H]OT), [³H]AVP and [³H]P₄ were from New England Nuclear. Membranes were incubated at 30°C for 60 min; under these conditions, binding reached equilibrium for all ligands used.

Inositol phosphate production. Measurements have been described^{25,26}. Cells were preincubated with vehicle or increasing amounts of P₄ for 15 min at 37°C before adding ligands and incubating for an additional 10 min.

[Ca_i] measurements. [Ca_i] was determined by dual-wavelength spectrometry^{16,25} using $0.7-1 \times 10^6$ Fura-2 loaded cells per experiment. Changes in [Ca_i] were quantitatively assessed by computerized integration of the increase over baseline during a 240-s stimulation period. Results are the mean $\pm$ s.e. of triplicate determinations; all experiments were done at least three times.

Received 27 November 1997; accepted 29 January 1998.

- Csapo, A. I., Puri, C. P. & Tarro, S. Relationship between timing of ovariectomy and maintenance of pregnancy in the guinea-pig. *Prostaglandins* **22**, 131-140 (1981).
- Pepe, G. J. & Albrecht, E. D. Actions of placental and fetal adrenal steroid hormones in primate pregnancy. *Endocr. Rev.* **16**, 608-648 (1995).
- Garfield, R. E., Puri, C. P. & Csapo, A. I. Endocrine, structural, and functional changes in the uterus during premature labor. *Am. J. Obstet. Gynecol.* **142**, 21-27 (1982).
- Fuchs, A. R., Periyasamy, S., Alexandrova, M. & Soloff, M. S. Correlation between oxytocin receptor concentration and responsiveness to oxytocin in pregnant rat myometrium: effects of ovarian steroids. *Endocrinology* **113**, 742-749 (1983).
- Soloff, M. S. *et al.* Regulation of oxytocin receptor concentration in rat uterine explants by estrogen and progesterone. *Can. J. Biochem. Cell Biol.* **61**, 625-630 (1983).
- Tsai, M. J. & O'Malley, B. W. Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annu. Rev. Biochem.* **63**, 451-486 (1994).
- Larcher, A. *et al.* Oxytocin receptor gene expression in the rat uterus during pregnancy and the estrous cycle and in response to gonadal steroid treatment. *Endocrinology* **136**, 5350-5356 (1995).
- Elands, J. *et al.* [¹²⁵I]-labelled d(CH₂)₂[Tyr(Me)]₂, Thr₄, Tyr-NH₂(9)]OVT: a selective oxytocin receptor ligand. *Eur. J. Pharmacol.* **147**, 197-207 (1988).
- Pepe, G. J. & Rothchild, I. A comparative study of serum progesterone levels in pregnancy and in various types of pseudopregnancy in the rat. *Endocrinology* **95**, 275-279 (1974).
- Schmidt, A. *et al.* A radioiodinated linear vasopressin antagonist: a ligand with high affinity and specificity for V_{1a} receptors. *FEBS Lett.* **282**, 77-81 (1991).
- Zingg, H. H. in *Bailliere's Clinical Endocrinology and Metabolism* (eds Sheppard, M. C. & Franklyn, J. A.) 75-96 (Bailliere Tindall, London, 1996).
- Burgisser, E., De Lean, A. & Lefkowitz, R. J. Reciprocal modulation of agonist and antagonist binding to muscarinic cholinergic receptor by guanine nucleotide. *Proc. Natl Acad. Sci. USA* **79**, 1732-1736 (1982).
- Green, R. D. Reciprocal modulation of agonist and antagonist binding to inhibitory adenosine receptors by 5'-guanylylimidodiphosphate and monovalent cations. *J. Neurosci.* **4**, 2472-2476 (1984).
- Sundaram, H., Newman-Tancredi, A. & Strange, P. G. Characterization of recombinant human serotonin 5HT_{1A} receptors expressed in Chinese hamster ovary cells. *Biochem. Pharmacol.* **45**, 1003-1009 (1993).
- Westphal, R. S. & Sanders-Bush, E. Reciprocal binding properties of 5-hydroxytryptamine type 2C receptor agonists and inverse agonists. *Mol. Pharmacol.* **46**, 937-942 (1994).
- Grazzini, E. *et al.* Membrane-delimited G protein-mediated coupling between V_{1a} vasopressin receptor and dihydropyridine binding sites in rat glomerulosa cells. *Mol. Pharmacol.* **50**, 1273-1283 (1996).
- Wassermann, W. J., Pinto, L. H., O'Connor, C. M. & Smith, L. D. Progesterone induces a rapid increase in [Ca²⁺]_i of *Xenopus laevis* oocytes. *Proc. Natl Acad. Sci. USA* **77**, 1534-1536 (1980).
- Majewska, M. D., Harrison, N. L., Schwartz, R. D., Barker, J. L. & Paul, S. M. Steroid hormone metabolites are barbiturate-like modulators of the GABA receptor. *Science* **232**, 1004-1007 (1986).
- Wong, M. & Moss, R. L. Patch-clamp analysis of direct steroid modulation of glutamate receptor-channels. *J. Neuroendocrinol.* **6**, 347-355 (1994).
- Valera, S., Ballivet, M. & Bertrand, D. Progesterone modulates a neuronal nicotinic acetylcholine receptor. *Proc. Natl Acad. Sci. USA* **89**, 9949-9953 (1992).
- Blackmore, P. F., Fisher, J. F., Spilman, C. H. & Bleasdale, J. E. Unusual steroid specificity of the cell surface progesterone receptor on human sperm. *Mol. Pharmacol.* **49**, 727-739 (1996).
- Schumacher, M., Coirini, H., Frankfurt, M. & McEwen, B. S. Localized actions of progesterone in hypothalamus involve oxytocin. *Proc. Natl Acad. Sci. USA* **86**, 6798-6801 (1989).
- Kimura, T., Tanizawa, O., Mori, K., Brownstein, M. J. & Okayama, H. Structure and expression of a human oxytocin receptor. *Nature* **356**, 526-529 (1992).
- Chini, B. *et al.* Tyr115 is the key residue for determining agonist selectivity in the V_{1a} vasopressin receptor. *EMBO J.* **14**, 2176-2182 (1995).
- Grazzini, E. *et al.* Dual effects of fluoro-aluminate on activation of calcium influx and inhibition of agonist-induced calcium mobilization in rat glomerulosa cells. *Cell Calcium* **19**, 29-41 (1996).
- Guillon, G. *et al.* Vasopressin stimulates steroid secretion in human adrenal glands: comparison with angiotensin-II effect. *Endocrinology* **136**, 1285-1295 (1995).
- Jeng, Y. J. *et al.* Molecular cloning and functional characterization of the oxytocin receptor from a rat pancreatic cell line (RINm5F). *Neuropeptides* **30**, 557-65 (1996).

Acknowledgements. We thank S. Jard, M. Bouvier and A. DeLean for discussion and for comments, N. Gallo-Payet and D. Payet for help with calcium measurements, J. Neculcea for technical help and E. Monaco for secretarial assistance. This research was supported from grants by the MRC of Canada. E.G. is supported by an MRC Fellowship award.

Correspondence and requests for materials should be addressed to H.H.Z. (e-mail: zingg@rvhri.lan.mcgill.ca).

Fatty acyl-CoA thioesters are ligands of hepatic nuclear factor-4 α

Rachel Hertz, Judith Magenheimer, Inna Berman & Jacob Bar-Tana

Department of Human Nutrition and Metabolism, Faculty of Medicine, Hebrew University, PO Box 12272, Jerusalem

Dietary fatty acids specifically modulate the onset and progression of various diseases, including cancer^{1,2}, atherogenesis³, hyperlipidaemia⁴, insulin resistance⁵ and hypertension⁶, as well as blood coagulability and fibrinolytic defects⁷; their effects depend on their chain length and degree of saturation. Hepatocyte nuclear factor-4 α (ref. 8) (HNF-4 α) is an orphan transcription factor of the superfamily of nuclear receptors and controls the expression of genes (reviewed in ref. 9) that govern the pathogenesis and course of some of these diseases. Here we show that long-chain fatty acids directly modulate the transcriptional activity of HNF-4 α by binding as their acyl-CoA thioesters to the ligand-binding domain of HNF-4 α . This binding may shift the oligomeric-dimeric equilibrium of HNF-4 α or may modulate the affinity of HNF-4 α for its cognate promoter element, resulting in either activation or inhibition of HNF-4 α transcriptional activity as a function of chain length and the degree of saturation of the fatty acyl-CoA ligands. In addition to their roles as substrates to yield energy, as an energy store, or as constituents of membrane phospholipids, dietary fatty acids may affect the course of a disease by modulating the expression of HNF-4 α -controlled genes.

Binding of fatty acyl-CoAs to HNF-4 α was verified using either the ligand-binding domain (LBD) of HNF-4 α fused to glutathione-S-transferase (GST) (GST-HNF-4 α (LBD)), or full-length HNF-4 α tagged by six histidine residues (His-HNF-4 α) (Fig. 1a). Binding of palmitoyl-CoA ((C16:0)-CoA) was saturable, having a K_d of 2.6 μ M (Fig. 1a, b) (range 1.2-3.4 μ M). This binding affinity conforms with liver cytosolic unbound long-chain fatty acyl-CoA, estimated at 1-4 μ M (ref. 10). (C16:0)-CoA binding at saturation approached a ratio of 1 mole acyl-CoA:1 mole HNF-4 α (Fig. 1b), assuming a single ligand-binding site per receptor in accordance with the single binding constant indicated by the binding curve. Binding to HNF-4 α was specific for the acyl-CoA, whereas the free fatty acid or free CoA was inactive. Binding of fatty acyl-CoAs of variable chain length and degrees of saturation was studied by using competing radiolabelled (C16:0)-CoA. Binding was not seen with saturated acyl-CoAs with chains of fewer than 12 carbon atoms. Binding of fatty acyl-CoAs with chains of more than 12 carbon atoms was not substantially affected by chain length or degree of saturation (Fig. 1c).

We determined whether the binding of acyl-CoAs to HNF-4 α was specific by analysing the putative binding of (C16:0)-CoA to peroxisome proliferator-activated receptor- α (PPAR α) or retinoic acid X receptor- α (RXR α). PPAR α was selected because it is activated by long-chain fatty acids¹¹⁻¹³, and RXR α in light of its homology with HNF-4 α (ref. 9). In contrast to the high-affinity, saturable binding of (C16:0)-CoA to HNF-4 α , binding to PPAR α or RXR α was of low affinity and non-saturable, indicating nonspecificity of binding (Fig. 1b).

The effect of HNF-4 α ligands on HNF-4 α -mediated gene transcription was evaluated in COS-7 cells that were cotransfected with an expression vector for HNF-4 α and with a chloramphenicol acetyltransferase (CAT) reporter plasmid enhanced by three C3P copies of the human apo CIII gene promoter sequence (nucleotides

–86 to –66) (ref. 14). Transfected cells were incubated with free fatty acids representing HNF-4 α prolignands of variable chain lengths and degrees of saturation. CAT expression was activated sevenfold by HNF-4 α in the absence of added fatty acids. Adding unsaturated long-chain fatty acids within their concentration range in plasma resulted in dose-dependent suppression of HNF-4 α transcriptional activity, with maximal effect observed with added (C18:3, ω -3) or (C20:5, ω -3) acids (Fig. 2a). The effect of saturated fatty acids was dependent on chain length: (C16:0) acid activated whereas (C18:0) acid suppressed HNF-4 α transcriptional activity. Saturated fatty acids shorter than C16 were inactive. Modulation of HNF-4 α activity by its prolignands was specific as, in agreement with previous reports^{11–13} and in contrast to results obtained with HNF-4 α , transfected murine PPAR α was shown to be activated by 0–200 μ M of either (C18:3, ω -3) or (C18:0) acids. HNF-4 α prolignands may serve as agonists or antagonists of HNF-4 α transcriptional

activity depending on their chain length and degree of saturation. Transcriptional activation in the absence of added exogenous prolignands could reflect an intrinsic activity of the unliganded receptor or result from binding of an endogenous activating acyl-CoA.

To determine whether there was a specific requirement for intracellular acyl-CoA in transfection studies, we studied the effects of representative agonistic or antagonistic HNF-4 α prolignands in cells cotransfected with long-chain fatty acyl-CoA synthase¹⁵ or thioesterase I (ref. 16) (Fig. 2b). Suppression of HNF-4 α transcriptional activity by (C18:3, ω -3) acid was more pronounced and occurred at lower concentrations of the free acid in cells cotransfected with long-chain fatty acyl-CoA synthase, but occurred at higher concentrations of the free acid in cells cotransfected with long-chain acyl-CoA thioesterase. Thus, the inhibitory effect of (C18:3, ω -3) acid was limited by conversion of the acid to the

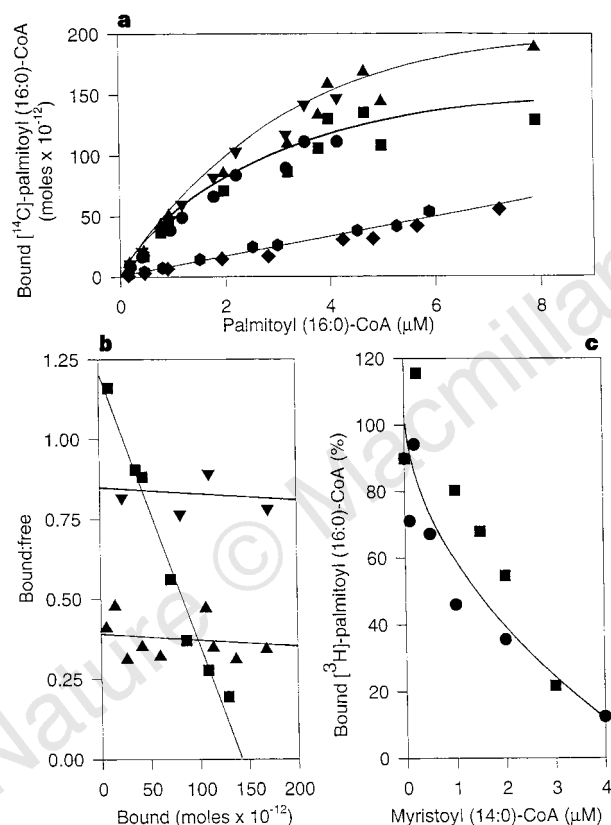


Figure 1 Long-chain fatty acyl-CoAs are ligands for HNF-4 α . **a**, Saturation binding curve for (C16:0)-CoA. Binding to HNF-4 α was determined by incubating His-HNF-4 α (140 pmol) (triangles) or GST-HNF-4 α (LBD) (140 pmol) (inverted triangles) with 4.3×10^{-3} μ Ci [14 C](C16:0)-CoA (57 mCi mmol $^{-1}$) and with increasing amounts of nonlabelled (C16:0)-CoA as indicated. Nonspecific binding is determined by (C16:0)-CoA binding to GST (140 pmol) (hexagons, bottom line) or carbonic anhydrase (110 pmol) (diamond). Specific binding of (C16:0)-CoA to His-HNF-4 α (squares) or to GST-HNF-4 α (LBD) (circles) is evaluated by subtracting the amount of nonspecific binding from the amount of total binding. **b**, Scatchard analysis of [14 C](C16:0)-CoA binding to His-HNF-4 α (squares) as compared with nonspecific (C16:0)-CoA binding to murine PPAR α (140 pmol) (inverted triangles) or human RXR α (140 pmol) (triangles). Conditions as in **a**. **c**, Competition for binding to HNF-4 α by (C14:0)-CoA. GST-HNF-4 α (LBD) (95.5 pmol) (circles) or His-HNF-4 α (140 pmol) (squares) is incubated with 8 nM [3 H](C16:0)-CoA (60 Ci mmol $^{-1}$) and with increasing amounts of nonlabelled (C14:0)-CoA as indicated. The y-axis indicates the percentage of radiolabelled [3 H](C16:0)-CoA in the bound fraction. 100% binding represents 0.3 pmol of [3 H](C16:0)-CoA. EC $_{50}$ (effective concentration resulting in 50% specific competition under the conditions employed) amounts to 1.4 μ M (range 1.2–1.5 μ M) of (C14:0)-CoA. EC $_{50}$ for other fatty acyl-CoAs are as follows: dodecanoyl(C12:0)-CoA, 2.3 μ M (range 2.1–2.4 μ M); palmitoyl(C16:0)-CoA, 2.6 μ M (range 1.3–3.4 μ M); stearoyl(C18:0)-CoA, 2.7 μ M (range 2.1–3.3 μ M); oleoyl(C18:1)-CoA, 1.4 μ M (range 1.0–1.8 μ M); linoleoyl(C18:2)-CoA, 1.9 μ M (range 1.5–2.3 μ M); linolenoyl(C18:3, ω -3)-CoA, 2.9 μ M (range 2.9–3.8 μ M); eicosapentaenoyl(C20:5, ω -3)-CoA, 0.6 μ M (range 0.5–0.7 μ M); docosahexaenoyl(C22:6, ω -3)-CoA, 1.6 μ M (range 0.6–2.7 μ M).

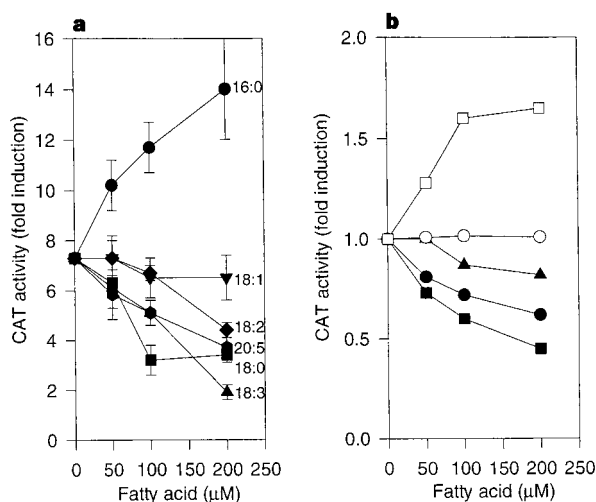


Figure 2 Modulation of HNF-4 α activity by long-chain fatty acids in transient transfection. **a**, The effects of fatty acid prolignands of HNF-4 α agonists and antagonists. Fold induction refers to CAT activity in cells transfected with pSG5-HNF-4 α as compared with cells transfected with pSG5. The mean $\pm$ s.e.m. is shown for 3–4 independent experiments for each fatty acid. **b**, HNF-4 α suppression by (C18:3, ω -3) acid (filled symbols) and HNF-4 α activation by (C14:0) acid (open symbols) in the absence (filled and open circles) and in response to (squares and triangles) cotransfected expression vectors encoding long-chain acyl-CoA synthase¹⁵ (2 μ g) (squares) or thioesterase I (ref. 16) (2 μ g) (triangles). Fold induction refers to CAT activity in cells incubated with the fatty acids, normalized to the activity in cells incubated in the absence of fatty acids. A representative experiment out of three similar experiments is shown.

respective acyl-CoA. Similarly, activation of HNF-4 α by (C14:0) acid was seen in cells enriched with acyl-CoA synthase but not in the absence of cotransfected acyl-CoA synthase (Fig. 2b), indicating that the endogenous availability and metabolic stability of the acyl-CoA are prerequisites for transcriptional modulation by HNF-4 α proligands. Activation of HNF-4 α by 100 μ M of (C16:0) acid was increased by 40% (range 16–75%) by cotransfected long-chain fatty acyl-CoA synthase or 10% (range 0–14%) decreased by fatty acyl-CoA thioesterase. Furthermore, in agreement with previous reports¹⁵ and in contrast to the response of HNF-4 α (Fig. 2b), activation of PPAR α by (C18:3, ω -3) acid was inhibited by cotransfected acyl-CoA synthase, thus indicating that the requirement for acyl-CoA was specific to HNF-4 α .

We studied the direct effects of representative agonistic and antagonistic HNF-4 α ligands by comparing the *in vitro* transcription rates of a test template, consisting of a guanine-free cassette enhanced by apo CIII C3P (ref. 14) as compared with internal control template lacking an HNF-4 α enhancer (Fig. 3). HNF-4 α -induced transcription of the test template was activated by added (C16:0)-CoA and inhibited by (C18:0)-CoA, concurring with the effect exerted by the respective HNF-4 α proligands in transfection studies (Fig. 2a).

We studied the binding of recombinant HNF-4 α dimer to its cognate C3P promoter (ref. 14) as a function of the type of added long-chain fatty acyl-CoAs using a gel mobility-shift assay. C3P binding was activated by saturated fatty acyl-CoAs with chain lengths of 14 and 16 carbon atoms (Fig. 4a) but was inhibited by (C18:0)-CoA and (C18:3, ω -3)-CoA (Fig. 4b). When amounts of HNF-4 α were limited, C3P binding was dependent on concentrations of (C14:0)-CoA (Fig. 4c) within the range of concentrations required for ligand binding to HNF-4 α (Fig. 1). C3P binding was not affected by the free acid of HNF-4 α acyl-CoA ligands. Activa-

tion of C3P binding by (C14:0)-CoA and inhibition by (C18:3, ω -3)-CoA were similarly observed using mammalian HNF-4 α translated *in vitro* in reticulocyte lysates (results not shown). However, there was no effect of these effectors on the mobility shift of *in vitro* translated PPAR α -RXR dimers. Fatty acyl-CoAs that bind to HNF-4 α may either activate or inhibit the binding of HNF-4 α to its cognate enhancer as a function of their chain length and degree of saturation, in agreement with the effects of the respective fatty acids and acyl-CoAs in transcription assays (Figs 2, 3). Binding of fatty acyl-CoAs to a defined transcription factor, resulting in the inhibition of its binding to its cognate DNA response element, has been reported for *Escherichia coli* FadR¹⁷.

Different long-chain fatty acyl-CoAs may affect the binding of HNF-4 α to its cognate enhancer by altering the availability of the HNF-4 α dimer and its intrinsic binding affinity¹⁸. We determined availability of the dimer by silver staining of recombinant His-HNF-4 α , which was subjected to electrophoresis under conditions and ligand concentrations similar to those used for gel mobility-shift studies. As for other receptors of the superfamily¹⁹, recombinant HNF-4 α preparations consist of oligomers, with a very low dimer content (Fig. 5). Adding (C14:0)-CoA or (C18:0)-CoA shifted the equilibrium to produce more HNF-4 α dimers or oligomers, respectively. These results concur with the respective agonistic or antagonistic effects of (C14:0)-CoA and (C18:0)-CoA in mobility-shift studies (Fig. 4).

The affinity of the HNF-4 α dimer for its cognate enhancer was determined by mobility-shift assays of radiolabelled C3P progressively diluted by non-labelled C3P. The apparent K_d for binding of C3P to His-HNF-4 α (40 ng) in the absence of added fatty acyl-CoA amounted to 1.5-ng oligonucleotide (range 1.2–2.0 ng) and remained unaffected by 10 μ M of added (C14:0)-CoA. However, adding (C18:3, ω -3)-CoA (10 μ M) resulted in a pronounced decrease in HNF-4 α binding to C3P (apparent K_d of 8.2–10 ng oligonucleotide), thus explaining why (C18:3, ω -3)-CoA is antagonistic even though it induces a relatively high content of HNF-4 α dimers (Fig. 5). Transcriptional modulation by HNF-4 α agonistic

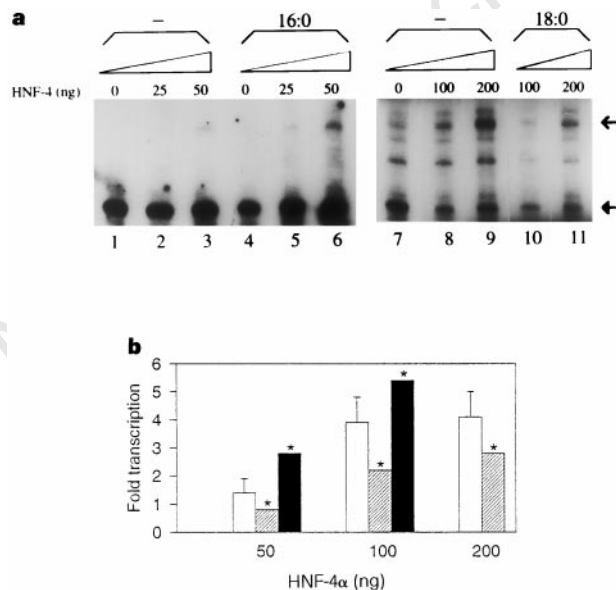


Figure 3 Modulation of HNF-4 α transcriptional activity by long-chain fatty acyl-CoAs *in vitro*. **a**, A representative experiment showing transcription of the test template by increasing concentrations of His-HNF-4 α in the absence (lanes 1–3, 7–9) or presence (lanes 4–6, 10, 11) of 10 μ M (C16:0)-CoA (lanes 4–6) or (C18:0)-CoA (lanes 10, 11). Correctly initiated transcripts of the test and control templates are denoted by the top and bottom arrows, respectively. **b**, HNF-4 α -induced *in vitro* transcription in the absence (open bars) or presence of 10 μ M each of (C16:0)-CoA (black bars) or (C18:0)-CoA (hatched bars). Fold transcription refers to the ratio of specific transcript produced from the test template to transcript produced from the control template, normalized to the ratio seen in the absence of HNF-4 α . The figure collates results from five independent experiments for each acyl-CoA. Asterisks indicate that these values are significant as compared with the respective value in the absence of added ligand.

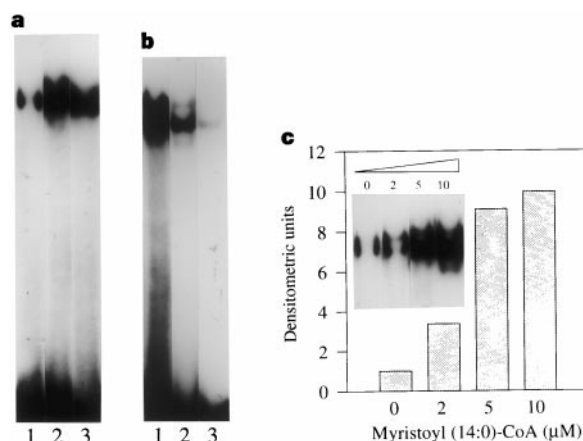


Figure 4 Fatty acyl-CoA ligands of HNF-4 α modulate its binding to its cognate enhancer DNA. **a**, His-HNF-4 α (14 ng) binding to C3P in the absence (lane 1) or presence (lanes 2, 3) of 10 μ M each of (C14:0)-CoA (lane 2) or (C16:0)-CoA (lane 3). **b**, His-HNF-4 α (20 ng) binding to C3P in the absence (lane 1) or presence (lanes 2, 3) of 10 μ M each of (C18:0)-CoA (lane 2) or (C18:3, ω -3)-CoA (lane 3). **c**, Activation of binding of His-HNF-4 α (14 ng) to C3P by increasing concentrations of (C14:0)-CoA. The gel section containing radiolabelled C3P bound to His-HNF-4 α dimers is shown.

or antagonistic acyl-CoA ligands may result from two apparently independent ligand-induced effects, namely, shifting the HNF-4 α oligomeric-dimeric equilibrium (for example, by (C14:0)-CoA, and (C18:0)-CoA) or affecting the intrinsic affinity of the HNF-4 α dimer for its cognate enhancer (for example, by (C18:3, ω -3)-CoA).

Our transfection and DNA-binding studies indicate that agonistic ligands of HNF-4 α may include saturated fatty acyl-CoAs with chains of 14 and 16 carbon atoms, whereas antagonistic ligands include some ω -3 and ω -6 polyunsaturated fatty acyl-CoAs and saturated (C18:0)-CoA. The overall effect exerted by HNF-4 α ligands on HNF-4 α -controlled genes may depend on the prevailing composition of nuclear acyl-CoAs within a specific cell type, as affected by the dietary intake of respective fatty acids, the availability of each fatty acid for CoA-thioesterification, and the metabolic stability of acyl-CoAs to hydrolysis, esterification or oxidation. Mild agonists of HNF-4 α may induce an apparent inhibition of HNF-4 α activity if replacing, or if competing with, an endogenous potent agonist. Also, as HNF-4 α may be modulated by phosphorylation²⁰, the effect of a certain fatty acid may reflect direct and indirect kinase-mediated activity. The effects may further depend on additional nuclear factors as well as putative acyl-CoA proligands other than fatty acids (for example, prostaglandins, and leukotrienes). In particular, as HNF-4 α and PPAR share similar DR-1 consensus sequences⁹, and as PPAR α is activated by long-chain free fatty acids rather than by their CoA thioesters^{11–13,15}, the effect exerted by a certain fatty acyl-CoA and mediated by HNF-4 α could be either similar to or antagonized by PPAR α which is activated by the respective free acid, resulting in synergistic or antagonist crosstalk between HNF-4 α and PPAR α .

The agonistic/antagonistic profile of fatty acyl-CoA ligands of HNF-4 α may offer a molecular basis for effects of dietary fatty acids *in vivo* in relation to some of the genes regulated by HNF-4 α . For example, an increase in levels of plasma very-low-density lipoprotein (VLDL)-, low-density lipoprotein (LDL)- and high-density lipoprotein (HDL)-cholesterol and their apoprotein constituents is induced by dietary saturated fatty acids of chains C12–C16 (ref. 4); saturated (C18:0) acid⁴ and polyunsaturated fatty acids⁴ have a lipid-lowering effect; an increase in blood coagulability is induced by dietary saturated C12–C16 fatty acids⁷; and a decrease in coagulability is induced by (C18:0) acid²¹ or polyunsaturated fatty acids⁷. These effects concerned with modulating blood lipids and coagulability may be ascribed to transcriptional modulation by the respective acyl-CoAs of HNF-4 α -controlled genes encoding apolipoproteins (AI, AII, B, CIII) and vitamin-K-dependent coagulability factors (VII, IX, X) (ref. 9). These examples may further indicate that some other genes²² or diseases^{1–3,5,6} reported to be modulated by dietary fatty acids could be controlled by HNF-4 α . Mutations in HNF-4 α and HNF-1 α were indeed shown to underlie maturity-onset diabetes of the young 1 and 3, respectively²³. Modulation of transcription of HNF-4 α -controlled genes by fatty acids may offer therapeutic treatment based on altering the activity of HNF-4 α by xenobiotic amphipathic carboxylates, which may be

thioesterified endogenously into their respective CoA thioesters and act as HNF-4 α agonists or antagonists. □

Methods

Recombinant proteins. Rat HNF-4 α 1 complementary DNA (pLEN4S) (ref. 8) subcloned into pGEX-2T (Pharmacia) was cleaved with *Sma*I and *Acc*I and religated. The GST–HNF-4 α (LBD) plasmid was induced in *E. coli* BL21 (DE3) by 0.2 mM isopropyl- β -D-thiogalactoside (IPTG) for 60 min and the product was purified over glutathione–agarose beads (Sigma) to yield the GST–HNF-4 α (LBD) protein. This protein consists of amino acids 96–455 of wild-type HNF-4 α fused to GST. The full-length HNF-4 α cDNA cloned into 6His-pET11d (ref. 24) was expressed in *E. coli* BL21(DE3)pLysS and purified over nickel affinity resin to yield the His–HNF-4 α protein. Murine PPAR α and human RXR α cDNAs were cloned into 6His-pET11d and expressed in *E. coli* BL21(DE3)pLysS. The PPAR α –RXR α heterodimer was shown to bind C3P (refs 14, 25).

Ligand-binding assays. Recombinant GST–HNF-4 α (LBD) or His–HNF-4 α was incubated for 60 min at 22 °C with [³H](C16:0)-CoA (American Radiolabeled Chemicals) or [¹⁴C](C16:0)-CoA (Amersham) in 100 μ l of 10 mM phosphate buffer (pH 7.4). Free and HNF-4 α -bound (C16:0)-CoA were separated by Dowex-coated charcoal²⁶.

Transfection assays. COS-7 cells cotransfected for 6 h with (C3P)₃-TK-CAT (5 μ g) and with either pSG5–HNF-4 α (ref. 25) (0.025 μ g) or pSG5 (0.025 μ g) were cultured in serum-free medium¹⁵ containing acids complexed to albumin in a molar ratio of 6:1. pRSGAL (1 μ g) added to each calcium-phosphate precipitate was an internal control for transfection. We prepared (C3P)₃-TK-CAT by inserting the apo CIII C3P enhancer (5'-GCAGGTGACCTTTGCCCC AGCGCC-3') (ref. 14) flanked by the *Hind*III restriction site into pBLCAT2. The expression vector for thioesterase I was prepared by cloning the *Nco*I/*Eco*RI insert of pJLA502 (ref. 16) into CMV4 (ref. 15).

In vitro transcription assays. Reaction mixture contained 20 mM Hepes-KOH (pH 7.9), 5 mM MgCl₂, 60 mM KCl, 8% glycerol, 2 mM DTT, 1 mM 3'-O-methyl-GTP, 10 units of T1 RNase, 20 units of RNasin, 0.5 μ g sonicated salmon sperm DNA, and His–HNF-4 α and test ligand as indicated. The mixture was preincubated for 30 min at 22 °C, and 10 ng of pAdML200 control template²⁷ and 200 ng of test template were then added. The mixture was further preincubated for 10 min at 22 °C. Next, 40 μ g of HeLa nuclear extract²⁸ was added and the mixture was again preincubated for 30 min at 30 °C. 0.5 mM ATP, 0.5 mM CTP, 25 μ M UTP, and 10 μ Ci of [α -³²P]UTP (s.a. 800 Ci mol⁻¹, Amersham) were then added and the complete reaction mixture was incubated for 45 min at 30 °C in a final volume of 25 μ l. RNA products were extracted, separated on a 5% polyacrylamide sequencing gel and quantified by Phosphor-imager. The test DNA template was constructed by inserting into the pC₂AT19 plasmid an oligonucleotide, amplified by the polymerase chain reaction, consisting of (C3P)₃ (ref. 14) and having *Eco*RI and *Sst*I sites at the 5' and 3' ends, respectively. The resulting plasmid was cleaved with *Sph*I and *Sac*I and ligated to a synthetic oligonucleotide (5'-CGAGGTCCACTTCGCTAT ATATTCCCCGAGCT-3') containing sequences of the herpes simplex virus thymidine kinase promoter (nucleotides –41 to –29) and of the chicken ovalbumin promoter (nucleotides –33 to –21) (ref. 29).

Gel mobility-shift assays. Mobility shift of ³²P-labelled C3P (0.1 ng) by HNF-4 α was analysed as described²⁵ using 0.6 $\times$ TBE.

Received 14 August 1997; accepted 12 January 1998.

1. Cave, W. T. Jr. Dietary n-3 (omega-3) polyunsaturated fatty acid effects on animal tumorigenesis. *FASEB J.* **5**, 2160–2166 (1991).
2. Welsch, C. W. Relationship between dietary fat and experimental mammary tumorigenesis: a review and critique. *Cancer Res. Suppl.* **52**, 2040–2048 (1992).
3. Chin, J. P. Marine oils and cardiovascular reactivity. *Prost. Leuk. Essent. Fatty Acids* **50**, 211–222 (1994).
4. Grundy, S. M. & Denke, M. A. Dietary influence on serum lipids and lipoproteins. *J. Lipid Res.* **31**, 1149–1172 (1990).
5. Storlien, L. H. *et al.* Fish oil prevents insulin resistance induced by high-fat feeding in rats. *Science* **237**, 885–888 (1987).
6. Morris, M. C., Sacks, F. & Rosner, B. Does fish oil lower blood pressure? A meta-analysis of controlled trials. *Circulation* **88**, 523–533 (1993).
7. Hultin, M. B. Fibrinogen and factor VII as risk factors in vascular disease. *Prog. Hemostasis Thromb.* **10**, 215–241 (1991).
8. Sladek, F. M., Zhong, W. M., Lai, E. & Darnell, J. E. Jr. Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev.* **4**, 2353–2365 (1990).
9. Sladek, F. M. in *Liver Gene Expression* (eds Tronche, F. & Yaniv, M.) 207–230 (R. G. Landes, Austin, Texas, 1994).

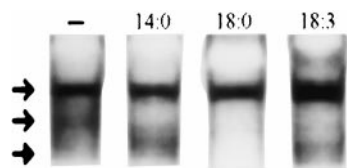


Figure 5 The oligomeric state of HNF-4 α is affected by long-chain fatty acyl-CoAs. His–HNF-4 α (400 ng) and fatty acyl-CoAs (10 μ M) were incubated and subjected to electrophoresis as in Fig. 4. Protein bands were visualized by silver stain. The top two arrows indicate the His–HNF-4 α dimer, as verified by a gel mobility-shift assay using radiolabelled C3P (Fig. 4) (refs 18, 24, 25). The bottom arrow indicates higher oligomers of His–HNF-4 α .

10. Gossett, R. E. *et al.* Acyl-CoA binding proteins: multiplicity and function. *Lipids* **31**, 895–918 (1996).
11. Forman, B. M., Chen, J. & Evans, R. M. Hypolipidemic drugs, polyunsaturated fatty acids, and eicosanoids are ligands for peroxisome proliferator-activated receptors α and δ . *Proc. Natl Acad. Sci. USA* **94**, 4312–4317 (1997).
12. Kliewer, S. A. *et al.* Fatty acids and eicosanoids regulate gene expression through direct interactions with peroxisome proliferator-activated receptors α and γ . *Proc. Natl Acad. Sci. USA* **94**, 4318–4323 (1997).
13. Krey, G. *et al.* Fatty acids, eicosanoids, and hypolipidemic agents identified as ligands of peroxisome proliferator activated receptors by coactivator dependent receptor ligand assay. *Mol. Endocrinol.* **11**, 779–791 (1997).
14. Leff, T., Reue, K., Melian, A., Culver, H. & Breslow, J. L. A regulatory element in the ApoCIII promoter that directs hepatic specific transcription binds to proteins in expressing and nonexpressing cell types. *J. Biol. Chem.* **264**, 16132–16137 (1989).
15. Hertz, R., Berman, I. & Bar-Tana, J. Transcriptional activation by amphipathic carboxylic peroxisomal proliferators is induced by the free acid rather than the acyl-CoA derivative. *Eur. J. Biochem.* **221**, 611–615 (1994).
16. Naggert, J., Witkowski, A., Wessa, B. & Smith, S. Expression in *Escherichia coli*, purification and characterization of two mammalian thioesterases involved in fatty acid synthesis. *Biochem. J.* **273**, 787–790 (1991).
17. DiRusso, C. C., Heimert, T. L. & Metzger, A. K. Characterization of FadR, a global transcriptional regulator of fatty acid metabolism in *Escherichia coli*. Interaction with the *fadB* promoter is prevented by long chain fatty acyl coenzyme A. *J. Biol. Chem.* **267**, 8685–8691 (1992).
18. Jiang, G. Q., Nepomuceno, L., Hopkins, K. & Sladek, F. M. Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. *Mol. Cell Biol.* **15**, 5131–5143 (1995).
19. Kersten, A., Kelleher, D., Chambon, P., Gronemeyer, H. & Noy, N. Retinoid X receptor alpha forms tetramers in solution. *Proc. Natl Acad. Sci. USA* **92**, 8645–8649 (1995).
20. Kistaki, E., Kistakis, N. T., Papadogeorgaki, E. & Talianidis, I. Recruitment of hepatocyte nuclear factor 4 into specific intranuclear compartments depends on tyrosine phosphorylation that affects its DNA-binding and transactivation potential. *Proc. Natl Acad. Sci. USA* **92**, 9876–9880 (1995).
21. Tholstrup, T., Marckmann, P., Jespersen, J. & Sandstrom, B. Fat high in stearic acid favorably affects blood lipids and factor VII coagulant activity in comparison with fats high in palmitic acid or high in myristic and lauric acids. *Am. J. Clin. Nutr.* **59**, 371–377 (1994).
22. Clarke, S. D. & Jump, D. B. Dietary polyunsaturated fatty acid regulation of gene transcription. *Annu. Rev. Nutr.* **14**, 83–98 (1994).
23. Yamagata, K. Mutations in the hepatocyte nuclear factor-4 α gene in maturity-onset diabetes of the young. *Nature* **384**, 458–460 (1996).
24. Malik, S. & Karathanasis, S. K. TFIIB-directed transcriptional activation by the orphan nuclear receptor hepatocyte nuclear factor 4. *Mol. Cell Biol.* **16**, 1824–1831 (1996).
25. Hertz, R., Bishara-Shieban, J. & Bar-Tana, J. Mode of action of peroxisome proliferators as hypolipidemic drugs. Suppression of apolipoprotein C-III. *J. Biol. Chem.* **270**, 13470–13475 (1995).
26. Dahlberg, E., Snochowski, M. & Gustafsson, J. A. Removal of hydrophobic compounds from biological fluids by a simple method. *Anal. Biochem.* **106**, 380–388 (1980).
27. Ori, A., Atzmoni, D., Haviv, I. & Shaul, Y. An NF1 motif plays a central role in hepatitis B virus enhancer. *Virology* **204**, 600–608 (1994).
28. Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11**, 1475–1489 (1983).
29. Klein-Hitpass, L. *et al.* The progesterone receptor stimulates cell-free transcription by enhancing the formation of a stable preinitiation complex. *Cell* **60**, 247–257 (1990).

Acknowledgements. We thank S. Malik for HNF-4 cloned into 6-HisPET11d, S. Smith for pJLA502, F. Sladek for pLEN4S plasmids and Y. Shaul for AdML 200-G-less cassette plasmids. This work was supported by the Israel National Science Foundation.

Correspondence and requests for materials should be addressed to J.B.-T. (e-mail: bartan@yam-suff.cc.huji.ac.il).

Dual function of the messenger RNA cap structure in poly(A)-tail-promoted translation in yeast

Thomas Preiss & Matthias W. Hentze

Gene Expression Programme, European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

The messenger RNA 3' poly(A) tail critically affects the initiation and control of translation in eukaryotes^{1–3}. By analogy to elements involved in transcription initiation, the poly(A) tail has been described as a 'translational enhancer' that enhances the 'translational promoter' activity of the mRNA 5'-cap structure^{3,4}. Elongation or shortening of the poly(A) tail regulates translation during development². Here we show, using cell-free and *in vivo* translation analyses in *Saccharomyces cerevisiae*, that the poly(A) tail can act as an independent 'translational promoter', delivering ribosomes to uncapped mRNAs even if their 5' end is blocked. When mRNAs compete for ribosome binding, neither the cap structure nor the poly(A) tail alone is enough to drive efficient

translation, but together they synergize and direct ribosome entry to the 5' end. The cap structure both promotes ribosome recruitment, together with the poly(A) tail, and tethers recruited ribosomes to the 5' end. Correct choice of translation initiation codons and the function of translational regulators acting on the 5' untranslated region are thus ensured by the functional interaction of the poly(A) tail with the cap structure.

In eukaryotes, the 5' cap structure (m⁷G(5')ppp(5')N) determines the quality and quantity of mRNA translation products. It strongly stimulates the frequency of translation initiation events on an mRNA and recruits the 40S ribosomal subunit to a site near the 5' end of the mRNA, thus enabling it to move along the mRNA in a 3' direction to identify the correct start site¹. An absolute requirement for the cap structure in cellular translation has been questioned by the discovery of specialized elements (internal ribosome entry sites, IRESs)⁵ in some viral and cellular mRNAs. These sites recruit ribosomes, even when the mRNA 5' end is blocked or uncapped. The IRES, too, affects quality and quantity of translation products in that initiation is directed to sites within or downstream of the IRES at rates reflecting its potency^{1,6}. Biochemical analyses of poly(A) tail function in translation became possible through the development of a cell-free translation system derived from *Saccharomyces cerevisiae*⁷; this system seems to recapitulate the strong synergism between cap structure and poly(A) tail that is observed *in vivo*⁸. Using these translation extracts, the poly(A) tail was found to stimulate recruitment of 40S ribosomal subunits⁹ and to allow translation of uncapped mRNA^{7,9}. Here we used this cell-free translation system and transfection of yeast spheroplasts with mRNA to understand the basic properties of the poly(A) tail in mediating translation initiation, and how these properties affect the quality and quantity of translation products together with the cap structure.

Reporter constructs encoding chloramphenicol acetyltransferase (CAT; Fig. 1a) were transcribed in four distinct versions: capped (c), polyadenylated (a), capped and polyadenylated (c-a) or neither capped nor polyadenylated (–). These mRNAs were translated in micrococcal-nuclease-treated yeast extract together with a capped and polyadenylated preprolactin internal control mRNA. Initial experiments with oligonucleotide target (OT)CAT mRNAs showed that the unmodified version is inefficiently translated in the yeast extract, whereas the poly(A) tail of OT.CATa mRNA drives significant CAT synthesis, with protein levels exceeding even those produced from cOT.CAT (see Supplementary information and Fig. 2a, lanes 1–4). This property of the poly(A) tail is largely due to stimulated translation, as parallel northern blot analyses showed a lack of corresponding differences in mRNA stability (measured half-lives: OT.CAT, 44 min; cOT.CAT, 44 min; OT.CATa, 58 min; cOT.CATa, 64 min). The observed relative stability of the uncapped mRNA in such assays^{7,9,10} may in part result from the presence of the 5' triphosphate group which, though translationally non-functional, distinguishes them from products of a physiological decapping reaction.

To test whether the translation of OT.CATa mRNA involves ribosome entry at the 5' end, we blocked defined regions of the mRNA using antisense 2'-O-allyl oligoribonucleotides. Such oligonucleotides inhibit cap-driven translation in rabbit reticulocyte lysate when they are hybridized to the mRNA within the 5' untranslated region (UTR), but do not inhibit translation when hybridized to the open reading frame (ORF) or the 3' UTR¹¹. As expected, translation of cOT.CAT mRNA is specifically blocked in the yeast translation reactions by each of seven different oligonucleotides directed against the 5' UTR. In contrast, translation of OT.CATa mRNA is only blocked by oligonucleotides targeted against the region at and just upstream of the initiation codon (Fig. 1b). This surprising result indicates that the poly(A) tail may drive translation not only independently of a cap structure but also when the 5' end is blocked. Such internal ribosome entry promoted

by the poly(A) tail may also explain the occurrence of smaller translation products (marked with asterisks in Fig. 2). Although variable in intensity, these products were routinely detected and are most prominent in 'a', diminished in 'c-a' and absent from 'c' mRNA translation reactions (see Supplementary information). A similar effect can be seen in experiments with luciferase mRNA constructs⁷.

To study 5'-end-independent ribosome entry and choice of 'internal' initiation codons systematically, we designed a new series of reporter constructs. uORF.CAT has a genetic, rather than a physical, block to 5'-end-initiated translation, because it contains the highly inhibitory upstream ORF (uORF) 4 from yeast *GCN4* mRNA¹². BIG.CAT mRNA encodes an amino-terminally extended polypeptide because of the mutation of the uORF stop codon to a lysine-encoding codon. This allows us to quantify the choice

between the in-frame start codons of uORF4 and CAT. As expected, cuORF.CAT mRNA does not yield detectable CAT protein and the translation product of cBIG.CAT mRNA (termed BIGCAT) is increased in size by 3K relative to the CAT protein (Fig. 2a, lanes 6, 10). The poly(A) tail bypasses the genetic block imposed by uORF4 of uORF.CATa mRNA (lane 7), and directs ~68% of initiation events to the downstream CAT initiation codon of BIG.CATa mRNA (lane 11). Comparison of the capped and polyadenylated with the polyadenylated versions of both constructs also shows that the cap structure redirects initiation to the start codon nearest the 5' end, resulting in decreased CAT synthesis from cuORF.CATa (Fig. 2a, lanes 3, 4, 7, 8) and a reversion of the ratio between CAT and BIGCAT protein from cBIG.CATa mRNA (Fig. 2a, lanes 11, 12).

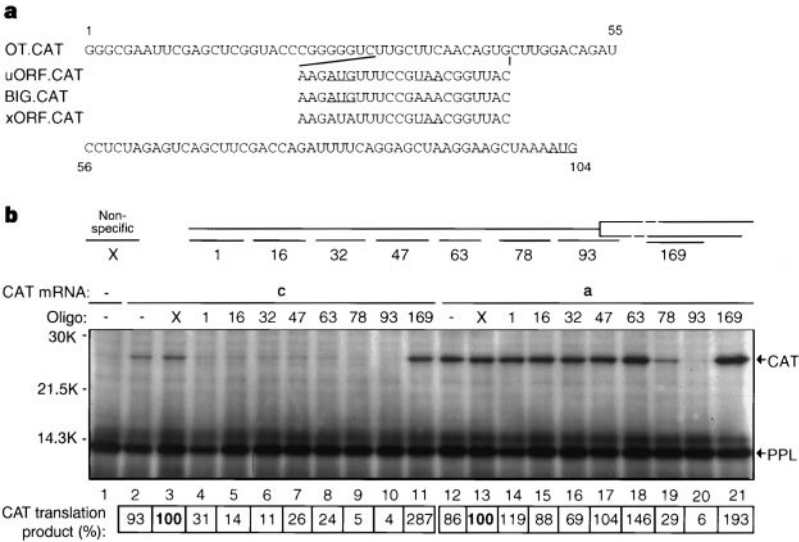


Figure 1 Poly(A)-tail-promoted translation is insensitive to a physically blocked mRNA 5' end. **a**, 5' UTR sequences of the various CAT mRNAs are shown with reference to OT.CAT mRNA (top and bottom, see Methods). Start and stop codons are underlined **b**, Preannealed mixtures of antisense 2'-O-allyl oligoribonucleotides (a diagram of the target positions is shown at the top of **b**) with 167 fmol OT.CAT mRNA and 8 fmol preprolactin (PPL) mRNA were translated in

nuclease-treated extract. Yield of CAT protein was normalized against the preprolactin control and expressed as the percentage of the result obtained with the non-specific oligonucleotide X (in bold) for each type of OT.CAT mRNA (see box below the figure). -, uncapped, non-adenylated CAT mRNA; a, polyadenylated CAT mRNA.

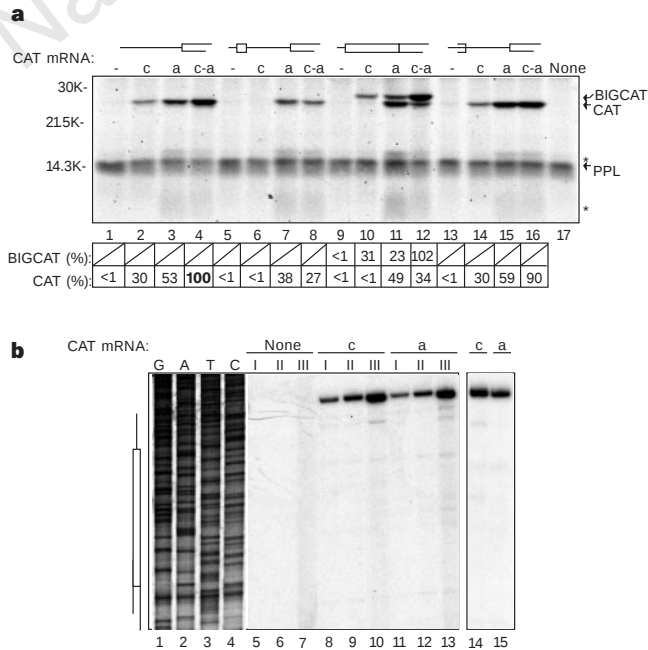


Figure 2 Initiation codon choice in cap- and poly(A)-tail-promoted translation. **a**, 100 fmol of the different CAT mRNAs and 2 fmol preprolactin (PPL) mRNA were translated in nuclease-treated extract: OT.CAT (lanes 1-4), uORF.CAT (lanes 5-8), BIG.CAT (lanes 9-12) and xORF.CAT mRNAs (lanes 13-16) were analysed (pictograms of the 5' UTR structure are shown above **a**). Normalized CAT and BIGCAT protein yields are given, below the gel, as a percentage of the value obtained with cOT.CATa mRNA (lane 4, in bold). Asterisks denote smaller CAT-mRNA-derived translation products. **b**, Translation reactions with either cBIG.CAT (lanes 8-10), BIG.CATa (lanes 11-13), or no added mRNA (lanes 5-7) were separated on linear 10-30% sucrose gradients and six fractions were collected. Aliquots of the isolated RNA from fractions sedimenting with 80S monosomes and faster (I-III) were analysed by reverse transcription using primer pxcat (see Methods). Lanes 14 and 15 are control reactions with input mRNAs. Sequencing reactions with the same primer act as size standards (lanes 1-4; initiation codons read C-A-T) lined up with a pictogram (at the side) of the BIG.CAT mRNA 5' region.

To be sure that the poly(A) tail promotes 'internal initiation', we studied the integrity of the translated mRNA. We isolated polysomal cBIG.CAT and BIG.CATa mRNAs and assessed their physical integrity in the critical region between the BIGCAT and CAT translation initiation codons by quantitative primer extension across this region (Fig. 2b). Both polysomal cBIG.CAT and BIG.CATa mRNAs were >90% intact (Fig. 2b, lanes 8–13) when ~68% of BIG.CATa mRNAs initiate translation at the CAT initiation codon (Fig. 2a, lane 11). This is inconsistent with the possibility that downstream translation products arise from mRNAs fragmented between the BIGCAT and CAT initiation codons. We followed the accumulation of BIGCAT and CAT polypeptides from BIG.CATa mRNAs over time (see Supplementary information). Accumulation of CAT protein should increase with incubation time if CAT protein were synthesized after nucleolytic removal of the BIGCAT initiation region from the mRNA in the assay, whereas production of BIGCAT protein should cease. We found a nearly constant ratio of BIGCAT:CAT proteins over 100 minutes of incubation (ratio at 10 min, 0.62; ratio at 100 min: 0.59), confirming our conclusion that the poly(A) tail promotes internal initiation from intact mRNAs.

These results indicate that, in this cell-free translation system, the poly(A) tail can promote translation independently of the cap and the mRNA 5' end and can direct initiation events to internal translation initiation sites. These properties concur with the view that the poly(A) tail promotes binding of the small ribosomal

subunit⁹, and cannot be adequately explained on the basis of a described⁴ stimulation of joining of the large subunit alone. The results also indicate that the cap structure can help tether ribosomes, recruited in a poly(A)-tail-mediated fashion, to the 5' end (Fig. 2a). The quantitative effects of the cap structure and the poly(A) tail on promoting translation in nuclease-treated extract are non-synergistic and, at best, additive (Fig. 2a). To study these unexpected properties in a living cell, we electroporated yeast spheroplasts with the same CAT reporter mRNAs. To analyse translation, we labelled cells with [³⁵S]methionine and examined CAT biosynthesis by quantitative immunoprecipitation. There are three major differences between the *in vivo* and *in vitro* data. First, the cap structure and the poly(A) tail are strongly synergistic *in vivo*. Second, although the poly(A) tail alone also promotes translation *in vivo*, its potency is substantially lower (Fig. 3). The cap structure alone is also rather inefficient in promoting translation *in vivo*. Third, choice of initiation codon in spheroplasts electroporated with the cBIG.CATa mRNA is strongly directed towards the 5' end (see Supplementary information; BIGCAT:CAT-protein ratio is ≥12).

Understanding these differences between the cell-free and *in vivo* systems could help us to determine the basis of the synergism between the cap structure and the poly(A) tail, as well as the means by which poly(A)-tail-copromoted translation is (largely) restricted to the 5' end. Rather than evoking the generation of novel, artefactual properties of the poly(A) tail in the course of extract preparation as an explanation, we considered mRNA competition

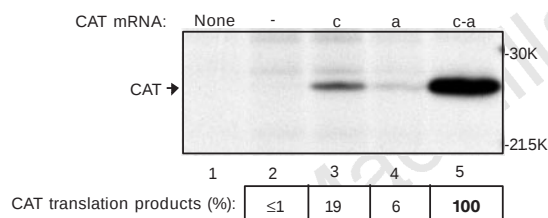


Figure 3 *In vivo* analysis of CAT-reporter mRNAs. OT.CAT mRNAs were electroporated into yeast spheroplasts and synthesis of CAT protein was studied by immunoprecipitation. Yields are given below the figure as a percentage of the value obtained with cOT.CATa mRNA (lane 5, in bold). The electroporated mRNAs were found to be similarly stable by northern analysis (half-life ≥100 min; see Supplementary information), consistent with previous reports^{8,19}. Equivalent results were obtained with xORF.CAT mRNAs (data not shown).

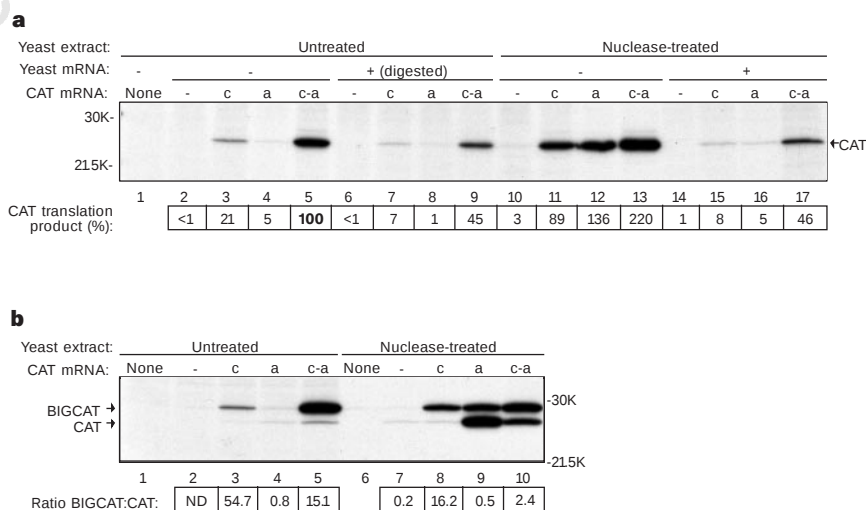


Figure 4 mRNA competition determines cooperativity between cap structure and poly(A) tail and enforces 5'-end-dependent ribosome recruitment. **a**, 167 fmol OT.CAT mRNA was translated in nuclease-treated (lanes 10–17) or mock-treated (lanes 1–9) extracts; CAT proteins were then immunoprecipitated. Some reactions were also supplemented with 400 ng oligo(dT)-cellulose-purified, yeast mRNA, either intact (lanes 14–17) or predigested with micrococcal nuclease (lanes 6–9). CAT protein yields are given below the figure as a percentage of the value obtained with cOT.CATa mRNA in untreated extract (lane 5, in bold). The

average stimulation of translation by nuclease treatment was calculated by comparing lanes 2–5 and 10–13, in this and several repeat experiments: OT.CAT, 36-fold increase; cOT.CAT, 5-fold increase; OT.CATa, 29-fold increase; cOT.CATa, 2-fold increase. Similarly strong changes were seen in luminescence assays (data not shown), with the previously used luciferase mRNAs^{7,9,10}. **b**, 167 fmol BIG.CAT mRNA was translated in nuclease-treated (lanes 6–10) or mock-treated (lanes 1–5) extracts. as in **a**. The molar ratios of BIGCAT to CAT protein are stated below the figure (averages from several experiments are given).

for initiation complex assembly as a possibility^{3,9,10,13}. We compared translation of OT.CAT mRNA *in vivo* (Fig. 3) with cell-free translation in mock-treated extracts (Fig. 4a, lanes 1–9) or in extracts from which endogenous competitor mRNAs had been removed by preincubation with micrococcal nuclease, as before (Fig. 4a, lanes 10–17). Because of the translation of endogenous mRNAs, CAT translation products were analysed by quantitative immunoprecipitation. Omission of micrococcal nuclease did not affect reporter mRNA stability (data not shown). In mock-treated extracts, the relative translation of the differently modified versions of OT.CAT mRNA is nearly identical to translation *in vivo*. This is important, as it shows that the functional characteristics of the cap structure and the poly(A) tail are preserved during extract preparation. Thus, the differences seen (see above) arise from a clearly defined step: the treatment with micrococcal nuclease. To address whether treatment with micrococcal nuclease causes its effects by removing endogenous competitor mRNAs and/or by other means, an amount of purified yeast mRNA appropriate to restore endogenous translation was substituted after micrococcal nuclease treatment of the extract (Fig. 4a, lanes 14–17). This substitution nearly recapitulates translation of OT.CAT mRNA in extracts without nuclease treatment and *in vivo* (Fig. 3). A small, but notable, extra effect results from the generation of yeast mRNA degradation products, as seen from the addition of predigested purified yeast mRNA to mock-treated extracts (Fig. 4a, lanes 2, 5, 6–9, 14–17). Furthermore, the use of the initiation codon nearest the 5' end of a capped and polyadenylated mRNA is enhanced in untreated extract (Fig. 4b). These results indicate that mRNA competition enforces the combined use of both the 'cap promoter' and the 'poly(A) tail promoter'. The results are also consistent with the finding that deadenylated mRNAs are translated well in yeast cells with the temperature-sensitive *pap1-1* mutation¹⁴, because of the absence of competition from capped and polyadenylated mRNAs.

These results indicate that the poly(A) tail can act as a translational promoter in its own right. Like the cap structure, the poly(a) tail alone is not potent enough to promote efficient translation when in competition with capped and polyadenylated mRNAs *in vivo* and *in vitro*. Our results thus provide functional evidence for a closed-loop model of translation initiation³, implying physical interactions between the mRNA 5' and 3' ends. The association of the poly(A)-binding protein Pab1p with the translation initiation factor eIF-4G suggests how this loop might be formed in yeast¹⁵. The cap structure and the poly(A) tail appear to function as near equal partners in determining the rate of mRNA translation. Furthermore, the cap structure also serves to direct the activity of the poly(A) tail to the 5' end. This is important for initiation-codon choice, the role of regulated poly(A) tail elongation and shortening^{2,16} and the function of translational regulatory elements acting on the 5' UTR. □

Methods

Plasmid constructs and *in vitro* transcription. Plasmid pOT.CAT is identical to pI19-CAT (ref. 11) except for four base changes in the 5' UTR (Fig. 1a). We generated puORE.CAT by replacing nucleotides 31–44 in pOT.CAT with sequences from uORF4 of the yeast *Gcn4* gene¹² in frame with the CAT coding region. Variants with a mutated start (ATG to ATA; pxORE.CAT) or stop (TAA to AAA; pBIG.CAT) codon were also generated. Each construct was also prepared carrying an (A)₉₈ segment in the 3' UTR, by ligating an *EcoRI/XbaI* fragment from pT3lucpA (ref. 7) into the *PstI* site. We confirmed all plasmid manipulations by sequencing. Transcripts were prepared as described (capping efficiency >95%) (ref. 17), omitting m⁷GpppG for uncapped mRNAs. Polyadenylated CAT mRNAs carry an (A)₉₈ tail followed by 25 nucleotides of linker sequence. mRNA concentration and integrity were assessed by trace-labelling and agarose-gel electrophoresis.

Cell-free translation. Extract preparations from yeast strains MBS (ref. 7) or YAS 1892 (*Matα ade2 leu2 trp1 his3 ura3 Mak10Δ pep4::HIS3*; a gift from A. Sachs; both yeast strains gave equivalent results) and translation reactions

were performed as described^{7,9} with some modifications. Extracts were pre-treated with micrococcal nuclease (75 or 150 units per ml, Pharmacia) for 5 min at 26 °C with continuous agitation. Translation reaction mixtures (15 µl) were incubated for 1 h at 23 °C in the presence of 0.37 MBq [³⁵S]methionine (37 TBq mmol⁻¹) and 2.5 mM magnesium acetate, and were digested with 4 µg RNase A at 37 °C for 20 min. For cotranslation, we chose the amounts of CAT-reporter and preprolactin-control mRNAs to avoid mRNA competition and to achieve comparable translation-product intensities. CAT, Nop1, U1A and the previously used luciferase^{7,9,10} mRNAs gave similar translation efficiencies in these assays, whereas the preprolactin-control mRNA was exceptionally active. Reactions shown in Figs 1, 2 were analysed by 15% SDS-PAGE and fluorography. CAT proteins shown in Fig. 4 were immunoprecipitated with anti-CAT antibodies (see below). Quantification was done with a Phosphor-Imager (Molecular Dynamics, ImageQuant software).

Sucrose-gradient centrifugation and primer-extension analysis. Translation reaction mixtures (50 µl/333 fmol CAT mRNA), using nuclease-treated extract, were incubated for 20 min as described, diluted with 100 µl ice-cold gradient buffer (30 mM Hepes/KOH, 170 mM potassium acetate, 2 mM magnesium acetate, 2 mM DTT) and fractionated on linear 10–30% sucrose gradients as described¹⁷, tracing the absorbance at 260 nm. Extracted RNA was subjected to primer extension analysis using Superscript II reverse transcriptase (Gibco). The primer pxcat (AGG CCG TAA TAT CCA GCT GAA C) hybridized 109 nucleotides downstream of the CAT initiation codon.

Electroporation of yeast spheroplasts. This was done as described¹⁸, with modifications: 280 µl of spheroplast suspension (2 × 10⁸ per ml) was supplemented with 10 µg CAT mRNA on ice, transferred immediately to an electroporation cuvette (diameter = 0.2 cm) and electroporated (at 800 V, 25 µF, 1000 Ω). After addition of 750 µl SD-methionine, 1 M sorbitol including and 3.7 MBq of [³⁵S]methionine (37 TBq mmol⁻¹), cells were agitated gently at 30 °C for 3.5 h. After centrifugation, the cell pellet was lysed in 1 ml solution containing 50 mM Tris/HCl pH 7.4, 300 mM NaCl, 1% Triton X-100, 0.1% BSA and 0.5 mM PMSE. Lysates (equal quantities of trichloroacetic-acid-precipitable radioactivity) were incubated with monoclonal anti-CAT antibody (a gift from C. Gorman) adsorbed to protein A sepharose, and analysed by 15% SDS-PAGE and fluorography. A second immunoprecipitation demonstrated the quantitative nature of this procedure.

Regional probes. 2'-O-allyl oligoribonucleotides were annealed to OT.CAT mRNA as described¹¹. Annealed mixtures containing mRNAs, 16 pmol oligonucleotide, and 190 ng *Escherichia coli* ribosomal RNA were added to translation reactions. Sequences of the oligonucleotides were (C with a number identifies complementarity to OT.CAT mRNA and the target nucleotide nearest the 5' end; X indicates an unrelated sequence): C1, AGC UCG AAU UCG CCC; C16, GAC CCC CGG GUA CCG; C32, AAG CZC UUU GZZ GCA G; C47, CUZ GZG GZU CUG UCC; C63, CUG GUC GAA GCU GAC; C78, UUZ GCU CCU GAA ZZU; C93, UUC UCC ZUU UUZ GCU UC; C169, TGA CUG AAA UGC CUC; X, CUA CAC GUC UAC CA (Z, 2,6-diaminopurine riboside; underlined G, non-complementary to OT.CAT mRNA; bold letters, deoxyribonucleotide).

Received 27 October 1997; accepted 9 January 1998.

1. Sachs, A. B., Sarnow, P. & Hentze, M. W. Starting at the beginning, middle, and end: translation initiation in eukaryotes. *Cell* **89**, 831–838 (1997).
2. Wickens, M., Kimble, J. & Strickland, S. in *Translational Control* (eds Hershey, J. W. B., Mathews, M. B. & Sonenberg, N.) 411–450 (Cold Spring Harbor Lab. Press, 1996).
3. Jacobson, A. in *Translational Control* (eds Hershey, J. W. B., Mathews, M. B. & Sonenberg, N.) 451–480 (Cold Spring Harbor Lab. Press, 1996).
4. Munroe, D. & Jacobson, A. mRNA poly(A) tail, a 3' enhancer of translational initiation. *Mol. Cell. Biol.* **10**, 3441–3455 (1990).
5. Pelletier, J. & Sonenberg, N. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334**, 320–325 (1988).
6. Jackson, R. J. in *Translational Control* (eds Hershey, J. W. B., Mathews, M. B. & Sonenberg, N.) 71–112 (Cold Spring Harbor Lab. Press, 1996).
7. Iizuka, N., Najita, L., Franzusoff, A. & Sarnow, P. Cap-dependent and cap-independent translation by internal initiation of mRNAs in cell extracts prepared from *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **14**, 7322–7330 (1994).
8. Gallie, D. R. The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. *Genes Dev.* **5**, 2108–2116 (1991).
9. Tarun, S. Z. Jr & Sachs, A. B. A common function for mRNA 5' and 3' ends in translation initiation in yeast. *Genes Dev.* **9**, 2997–3007 (1995).
10. Tarun, S. Z. Jr, Wells, S. E., Deardorff, J. A. & Sachs, A. B. Translation initiation factor eIF-4G mediates *in vitro* poly(A) tail-dependent translation. *Proc. Natl Acad. Sci. USA* **94**, 9046–9051 (1997).
11. Johansson, H. E., Belsham, G. J., Sproat, B. S. & Hentze, M. W. Target-specific arrest of mRNA translation by antisense 2'-O-alkyloligoribonucleotides. *Nucleic Acids Res.* **22**, 4591–4598 (1994).
12. Hinnebusch, A. G. in *Translational Control* (eds Hershey, J. W. B., Mathews, M. B. & Sonenberg, N.) 199–244 (Cold Spring Harbor Lab. Press, 1996).

13. Gallie, D. R. & Tanguay, R. Poly(A) binds to initiation factors and increases cap-dependent translation *in vitro*. *J. Biol. Chem.* **269**, 17166–17173 (1994).
14. Proweller, A. & Butler, J. S. Ribosome concentration contributes to discrimination against poly(A)-mRNA during translation initiation in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **272**, 6004–6010 (1997).
15. Tarun, S. Z. Jr. & Sachs, A. B. Association of the yeast poly(A) tail binding protein with translation initiation factor EIF-4G. *EMBO J.* **15**, 7168–7177 (1996).
16. Muckenthaler, M., Gunkel, N., Striebeck, R. & Hentze, M. W. Regulated poly(A) tail shortening in somatic cells mediated by cap-proximal translational repressor proteins and ribosome association. *RNA* **3**, 983–995 (1997).
17. Gray, N. K. & Hentze, M. W. Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs. *EMBO J.* **13**, 3882–3891 (1994).
18. Everett, J. G. & Gallie, D. R. RNA delivery in *Saccharomyces cerevisiae* using electroporation. *Yeast* **8**, 1007–1014 (1992).
19. Masson, D. C. *et al.* Decoying the Cap⁺ mRNA degradation system by a double-stranded RNA virus and poly(A)⁺ mRNA surveillance by a yeast antiviral system. *Mol. Cell. Biol.* **15**, 2763–2771 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank A. Sachs and P. Sarnow for reagents and stimulating discussions, and A. Ephrussi, A. Jenny, F. Stewart, and the members of the Hentze lab for comments on the manuscript. This work was supported by an EMBO fellowship to T.P.

Correspondence and requests for materials should be addressed to M.W.H. (e-mail: Hentze@EMBL-Heidelberg.de).

Interaction of polyadenylate-binding protein with the eIF4G homologue PAIP enhances translation

Andrew W. B. Craig, Ashkan Haghighat, Annie T. K. Yu & Nahum Sonenberg

Department of Biochemistry and McGill Cancer Centre, McGill University, Montreal, Quebec H3G 1Y6, Canada

In the initiation of translation in eukaryotes, binding of the small ribosomal subunit to the messenger RNA results from recognition of the 5' cap structure (m⁷GpppX) of the mRNA by the cap-binding complex eIF4F¹. eIF4F is itself a three-subunit complex comprising the cap-binding protein eIF4E², eIF4A, an ATP-dependent RNA helicase³, and eIF4G, which interacts with both eIF4A and eIF4E and enhances cap binding by eIF4E⁴. The mRNA 3' polyadenylate tail and the associated poly(A)-binding protein (PABP) also regulate translational initiation⁵, probably by interacting with the 5' end of the mRNA^{6,7}. In yeast^{8,9} and plants¹⁰, PABP interacts with eIF4G^{8,9} but no such interaction has been reported in mammalian cells. Here, we describe a new human PABP-interacting protein, PAIP-1, whose sequence is similar to the central portion of eIF4G and which interacts with eIF4A. Overexpression of PAIP-1 in COS-7 cells stimulates translation, perhaps by providing a physical link between the mRNA termini.

To identify proteins that interact with human PABP in a far-western assay¹¹, we constructed a recombinant baculovirus expressing an amino-terminal-tagged version of PABP. This Flag-epitope tag enabled the ~80K Flag-PABP fusion protein to be affinity-purified from insect cells. This was then labelled with ³²P using heart muscle kinase (HMK), and used to probe blots of purified translation-initiation factors. No interaction with known initiation factors, including eIF4G, was detected (data not shown), but a few polypeptides in a HeLa extract, including a 70K polypeptide, bound Flag-PABP in far-western assays (data not shown). A human placental complementary DNA expression library was therefore screened using the far-western technique, and a partial cDNA fused in-frame to β-galactosidase was identified. Several additional cDNAs were identified by using a DNA probe from this clone. The largest cDNA (2.8 kilobases (kb)) encoded a 480-amino-acid protein, PAIP-1 (for PABP-interacting protein-1). The predicted

amino-acid sequence of PAIP-1 (Fig. 1) shows significant similarity (25% identity and 39% similarity) with the central portion of human eIF4G¹² (amino acids 420 to 890). This portion of eIF4G contains one of two known eIF4A-binding regions¹³ and also binds to another initiation factor, eIF3 (refs 13, 14). The residues in eIF4G that are required for eIF4E binding¹⁵ are underlined; these are not present in PAIP-1. PAIP-1 also shows significant homology to the eIF4G-related protein p97 (ref. 16) (also called DAP-5 and NAT-1; refs 17, 18).

To test whether PABP and PAIP-1 interact *in vitro*, histidine-tagged (His-) PAIP-1 was incubated with poly(A)-Sephacrose or Flag monoclonal-antibody-coupled Sepharose, with or without Flag-PABP. Bound proteins were resolved by SDS-PAGE and detected by Coomassie brilliant blue staining (Fig. 2a). His-PAIP-1 migrated as a ~70K protein (lane 1; indicated by a dot), which is larger than its predicted relative molecular mass of 54K, probably because of its proline-rich N terminus. Purified Flag-PABP (Fig. 2a, lane 2; indicated by a dot) bound strongly to both poly(A)- and Flag antibody-coupled Sepharose (Fig. 2a, lanes 4 and 5, 7 and 8; >70% binding). In contrast, PAIP-1 did not itself bind to poly(A)- or Flag antibody-coupled Sepharose (Fig. 2a, lanes 3 and 6), but was recovered in the presence of PABP (Fig. 2a, lanes 5 and 8; a minor protein species of ~75K that co-purified with His-PAIP-1 (lane 1) was recovered with poly(A)- (lanes 3, 5) but not with Flag antibody-Sepharose (lanes 6, 8)). The amount of PAIP-1 recovered with PABP on poly(A)-Sepharose was ~20% (Fig. 2a, lane 5) and ~60% of the input with Flag antibody-coupled Sepharose (lane 8). These results show that purified PAIP-1 and PABP interact *in vitro*. This interaction is not dependent on polyadenylated RNA and probably not on any other RNA species, given that micrococcal nuclease treatment had no effect on binding in far-western and co-precipitation assays (data not shown).

We tested for an association between native PAIP-1 and PABP *in vivo* by co-immunoprecipitation assay with a HeLa S10 extract.

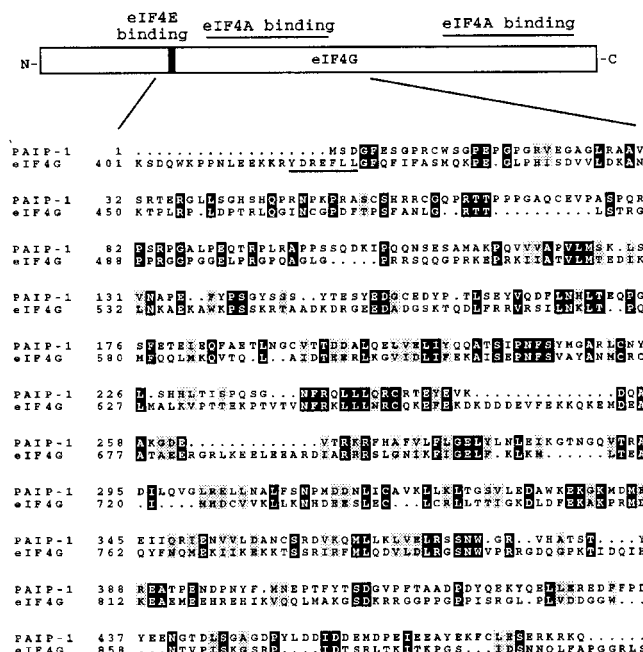


Figure 1 Identification of PAIP-1 and its homology with eIF4G. The eIF4E-binding site¹⁸ is underlined. Two eIF4A-binding sites within eIF4G¹³ are indicated. Alignment of the amino-acid sequences of PAIP-1 and eIF4G was carried out using PIMA multisequence alignment (Baylor College of Medicine Search Launcher). Identical amino-acid residues are boxed and conservative substitutions are shaded. Amino-acid numbers are shown on the left.

Anti-PAIP-1 and PABP antibodies were tested for crossreactivity by immunoprecipitation of recombinant His-PAIP-1 and His-PABP proteins, followed by western blotting with an antibody directed against the Xpress epitope at the N terminus of both proteins (see Methods). His-PAIP-1, but not PABP, was recovered with PAIP-1 antisera (Fig. 2b, lanes 1 and 2), whereas the PABP antiserum recovered only His-PABP (lanes 3 and 4); there is therefore no crossreactivity between PABP and PAIP-1 antisera. A HeLa S10 extract was next incubated with antibodies against affinity-purified glutathione-S-transferase (GST), PAIP-1 or PABP. Following precipitation with protein G-Sepharose, bound proteins were western blotted with antibodies against the proteins indicated, and the amount of protein precipitated was compared to the total (Fig. 2c, lane 1; 20% of the input was analysed). PABP was detected in both

the PAIP-1 and the PABP immunoprecipitates (Fig. 2c, top panel; compare lanes 3, 4 and 1; 12 and 40% recovery, respectively), but not in the GST immunoprecipitate (lane 2). PAIP-1 (Fig. 2c, lane 1, second panel) was recovered efficiently with PAIP-1 antisera (lane 3, 36% recovery; the 50K band in lanes 2 and 3 is the antibody heavy chain), and also immunoprecipitated with the PABP antibody (4% of input (lane 4), allowing for the small amount of PAIP-1 that immunoprecipitated with antibody against GST (lane 2)). Thus, PAIP-1 and PABP form a complex in HeLa extracts.

As PAIP-1 contains a region with significant homology to one of the two eIF4A-binding domains in eIF4G¹³, we also tested for co-immunoprecipitation of eIF4A with the PAIP-1 antiserum. eIF4A was recovered in the PAIP-1 immunoprecipitate (Fig. 2c, lane 3, third panel; 10% recovery) and to a lesser extent in the PABP

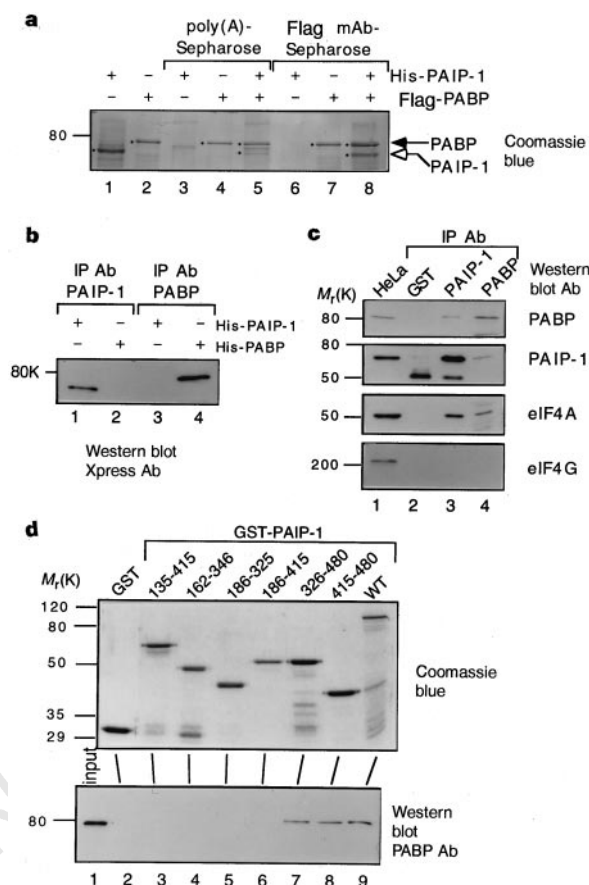


Figure 2 Interaction of PAIP-1 with PABP. **a**, PAIP-1 binds to poly(A)- and Flag monoclonal antibody (mAb)-coupled Sepharose in the presence of PABP. Coomassie brilliant blue staining of input His-PAIP-1 (lane 1), Flag-PABP (lane 2), and proteins bound to poly(A)-Sepharose (lanes 3-5) and to Flag mAb-Sepharose (lanes 6-8). **b**, PAIP-1 and PABP antisera do not crossreact. Purified His-PAIP-1 and His-PABP were immunoprecipitated with both PAIP-1- and PABP-specific antibodies followed by western blotting with anti-Xpress antibody. **c**, Co-immunoprecipitation of PABP and eIF4A with PAIP-1. HeLa S10 extracts were incubated with affinity-purified GST and PAIP-1 specific antibodies or with a PABP monoclonal antibody, and western-blotted with different antibodies against the proteins indicated to the right. **d**, Mapping of the PABP-binding site on PAIP-1. Coomassie brilliant blue staining (top panel) of purified GST (lane 2) and GST-PAIP-1 fusion proteins (lanes 3-9). Flag-PABP protein (lane 1, bottom panel) was preincubated with GST and GST-PAIP-1 fusion proteins, precipitated with glutathione-Sepharose beads and detected by western blotting. IP, immunoprecipitation; Ab, antibody.

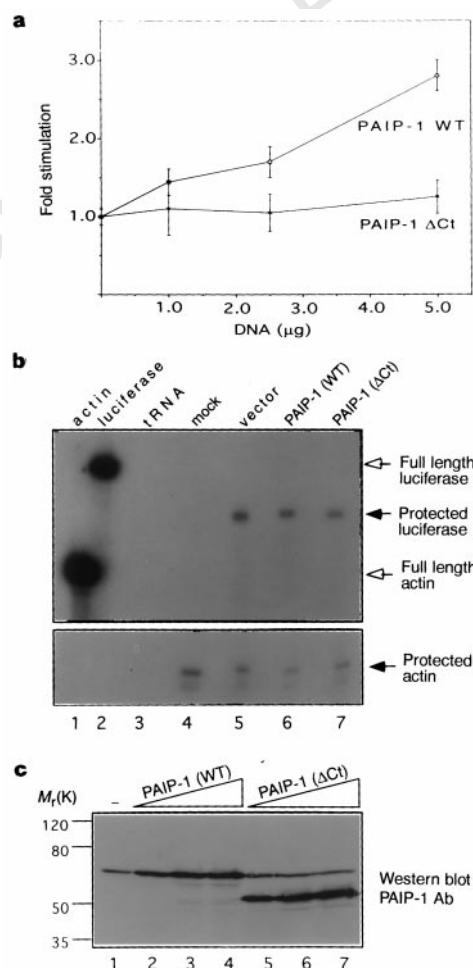


Figure 3 PAIP-1 overexpression enhances translation. COS-7 cells were transiently co-transfected with a luciferase reporter cDNA, and either pcDNA3 alone or the indicated amounts of pcDNA3-PAIP-1 (WT) or pcDNA3-PAIP-1(ΔC1) plasmids. **a**, Luciferase activity is shown; the mean value for six assays in three independent transfections is given, with the standard deviation as error bars. **b**, A representative RNAse protection assay. Antisense actin and luciferase probes (lanes 1 and 2) were tested against tRNA (lane 3), and RNA from mock (lane 4) or transfected cells (lanes 5-7), as described in Methods. The positions of the full-length probes and protected fragments are indicated on the right. **c**, Wild-type PAIP-1 and PAIP-1(ΔC1) proteins are expressed to similar levels in COS-7 cells. Protein extracts from transfected cells were subjected to SDS-10% PAGE followed by western blotting with PAIP-1 antisera. Relative molecular mass markers are shown on the left.

immunoprecipitate (lane 4, 3%), suggesting that a ternary complex between PAIP-1, PABP and eIF4A is formed. eIF4G was only detected in the input (Fig. 2c, bottom panel; lane 1) and not in any of the immunoprecipitates (lanes 2–4). The association of PAIP-1 with PABP and eIF4A, and the lack of interaction between PABP and eIF4G, was confirmed using a two-hybrid assay in yeast (data not shown).

To identify the PABP-binding site in PAIP-1, different parts of PAIP-1 were fused to GST and the proteins expressed in *Escherichia coli*. The purified proteins (Fig. 2d, top panel) were preincubated with Flag-PABP, precipitated with glutathione–Sepharose, and western blotted with an antibody against PABP (bottom panel). PABP was detected in the input material (Fig. 2d, lane 1), and in reactions containing fusion proteins retaining the C terminus of PAIP-1 (lanes 7–9). This portion of PAIP-1 (amino acids 415 to 480) is rich in acidic amino acids and is not conserved in eIF4G (Fig. 1). These results indicate that the C terminus of PAIP-1, comprising amino acids 415 to 480, is sufficient for interaction with PABP.

To investigate whether PAIP-1 participates in translation, we co-transfected COS-7 cells with vectors expressing either wild-type PAIP-1 or a carboxy-terminal deletion mutant lacking the PABP-binding site, PAIP-1(ΔC_t), together with luciferase reporter plasmid (Fig. 3). With increasing amounts of the wild-type PAIP-1 plasmid, there was a dose-dependent increase of up to 2.8-fold in luciferase activity compared to that for vector alone (Fig. 3a). Transfection with the PAIP-1(ΔC_t)-expressing plasmid caused no stimulation, suggesting that the C terminus of PAIP-1 is required for this effect. As this portion of PAIP-1 confers its ability to associate with PABP, this would suggest that translational stimulation requires the interaction of PAIP-1 with PABP. To exclude the possibility that the effects observed reflect differences in mRNA concentration, we did

RNAse protection assays on RNA extracted from parallel transfections using antisense actin and luciferase probes (Fig. 3b, lanes 1 and 2). As expected, there were no protected fragments for the transfer RNA control (Fig. 3b, lane 3), and only actin mRNA was detected in mock-transfected cells (lane 4, bottom panel). The luciferase/actin mRNA ratios were 1.2 ± 0.1 for wild type and 0.8 ± 0.1 for PAIP-1(ΔC_t), relative to vector (set at 1) for three experiments (Fig. 3b, lanes 5–7). These results indicate that there may be small differences in the amounts of luciferase mRNA, but these are not sufficient to explain the increase in luciferase activity. Overexpression of PAIP-1 therefore enhances translation in mammalian cells. Western blotting with PAIP-1 antiserum indicated that wild-type and PAIP-1(ΔC_t) proteins were comparably expressed (Fig. 3c). Overexpressed wild-type PAIP-1 co-migrated with a protein from COS cells that was detected with PAIP-1 antiserum (Fig. 3c; compare lane 1 with lanes 2–4) and is likely to be the simian homologue of PAIP-1. Because PAIP-1 appears to be present at ~6-fold lower levels than PABP (A.W.B.C., data not shown), its overexpression would be likely to result in more PAIP-1/PABP complex and to potentiate contacts with the 5' end of mRNA by interacting with eIF4A.

Of what significance is the interaction of PAIP-1 with both PABP and eIF4A? The eIF4A subunit of eIF4F is required for translation of all mRNAs, as dominant-negative mutants of eIF4A repress both cap-dependent and cap-independent translation¹⁹. It has been proposed that loading of eIF4A onto the 5' untranslated region (UTR) is followed by the unwinding of mRNA secondary structure, resulting in enhanced ribosome binding¹. The interaction between PAIP-1 and eIF4A could enable bridging to occur between the PAIP-1/PABP complex on the poly(A) tail and eIF4A bound to the 5' UTR (Fig. 4a). Such an interaction between the mRNA 5' and 3' ends might select only intact mRNAs (containing both cap and poly(A) tail) as templates for translation and may also protect the mRNA from degradation. Moreover, proximity of the mRNA ends may promote reinitiation of terminating ribosomes on the same mRNA, thereby enhancing translation. An early study on translation in rabbit reticulocyte lysates indicated that reinitiating ribosomes are less sensitive to inhibition by cap analogues than those participating in primary initiation²⁰. This is consistent with the utilization by a reinitiating ribosome of an alternative, cap-independent route onto the mRNA, such as that provided by a putative link formed between the mRNA termini (Fig. 4). It is possible that PAIP-1 links PABP to the 40S ribosome subunit, as predicted in yeast^{8,21}, through an interaction with eIF3, although we were unable to detect this (A.W.B.C., unpublished).

In yeast and in plants, the interaction between Pab1 and eIF4G provides a bridge between the cap and poly(A) tail^{8–10} (Fig. 4b). However, the lack of an observed interaction between human PABP and eIF4G (Fig. 2e; S. Wells and A. Sachs, personal communication) suggests that this mechanism^{7,8,22} might not operate in mammalian cells. We have recently identified a new human homologue of eIF4G, as well as a longer form of the original eIF4G²³, but their ability to interact with PABP needs to be determined. The bridging of mRNA termini in mammalian cells by an adaptor protein such as PAIP-1 may permit additional levels of regulation not required in yeast. For example, PAIP-1 is induced following T-cell activation and interacts with the T-cell-inducible iPABP²⁴ (A.W.B.C. and T. Lindsten, unpublished results). This is consistent with the idea that PAIP-1 is a positive effector of translation. Thus, the presence of PAIP-1 in animal cells may reflect an evolutionary advance, allowing higher eukaryotes to link PABP to eIF4A function without directly affecting the cap-binding process. □

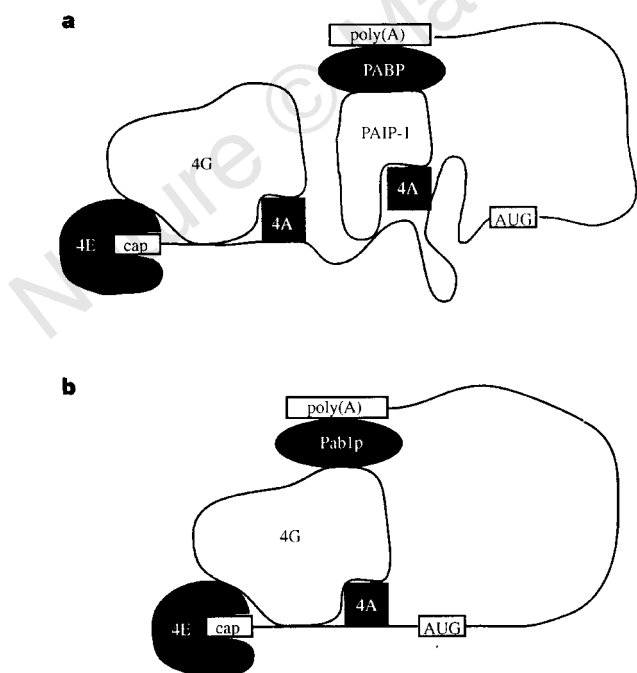


Figure 4 Proposed models for the bridging of 5' and 3' mRNA ends in eukaryotes. **a**, Association of PAIP-1 with eIF4A and PABP links the mRNA termini in mammalian cells. **b**, A direct interaction between PABP and eIF4G in yeast and plant cells (adapted from ref. 29). The position of the cap structure, initiator methionine codon (AUG), and poly(A) tail are boxed. eIF4E is shown binding the cap structure. eIF4G enhances binding of eIF4E to the cap by interaction with the mRNA⁴. eIF4A recycles through the cap-binding complex and is thought to act together with eIF4B in mRNA unwinding¹³.

Methods

Plasmids. A 1.7-kb clone, obtained by far-western screening of a human placental cDNA library with ³²P-Flag-HMK-PABP, was digested with *Sall* and subcloned into the *Sall* site of KS vector (Stratagene), forming the KS-S

plasmid. A larger clone (2.8 kb) containing the complete coding region for PAIP-1, was obtained by rescreeing the library, and was designated KS-S8. A 115-nucleotide truncation of the 5' UTR (S8Δ5') was generated by digestion of KS-S8 with *NarI/XbaI*, followed by ligation to SK (Stratagene) cut with *AccI/XbaI*.

A *KpnI/XbaI* fragment from SK-S8Δ5' was cloned into the vector pcDNA3 (Invitrogen) to allow the expression of PAIP-1 in mammalian cells. A C-terminal deletion construct [PAIP-1(ΔC)] was generated by digestion of SK-S8Δ5' with *KpnI/PvuII* (which truncates the coding region at amino acid 415), and ligated to pcDNA3 previously digested with *KpnI/EcoRV*. To generate GST-PAIP-1 fusion proteins, the *SalI* fragment from the KS-S plasmid (encoding amino acids 4–480) was ligated to pGexHMK¹¹. GST-PAIP-1(135–415) was generated by ligation of an *AseI/PvuII* fragment from KS-S to pGex3X (Pharmacia). GST-PAIP-1(162–346) was obtained by ligation of an *XmnI* fragment from KS-S to pGexHMK. GST-PAIP-1(186–325) and (186–415) fusions were generated by ligation of *BsaI/HincII* and *BsaI/PvuII* fragments to pGex3X. GST-PAIP-1(326–480) and (415–480) were generated by ligation of *HincII/ClaI* and *PvuII/ClaI* fragments from KS-S8 to pGexHMK and pGex3X.

Generation of recombinant baculoviruses. To generate a Flag-HMK¹¹ fusion of PABP, an *NcoI-BamHI* fragment of pHu73 (encoding human PABP²⁵) was subcloned into the *EcoRI* site of the baculovirus transfer vector pVL1392-FlagHMK²⁶ (Pharmingen). Recombinant baculovirus was generated with the BaculoGold expression system (Pharmingen). Flag-PABP was immunopurified using a commercial anti-FLAG mAb column (Kodak). PAIP-1 and PABP were expressed in Sf9 cells with six-histidine and Xpress epitope N-terminal tags (fused to the *SalI* fragment from clone S and the *NcoI/BamHI* fragment from pHu73 (ref. 25), respectively) using the pBlueBacHis2 baculovirus vector (Invitrogen), and purified on Ni²⁺-NTA resin (Qiagen).

Screening of a human cDNA expression library. A human placenta λgt11 cDNA expression library (kindly provided by M. Park) was plated (1 × 10⁶ phages), and transferred to IPTG-soaked nitrocellulose filters (Amersham). The far-western probe was generated by incubation of purified Flag-HMK-PABP (2 μg) with heart muscle kinase (10 U; Sigma) in the presence of [γ-³²P]-ATP (50 μCi) as described¹¹.

Isolation of RNA and RNase protection assay. Total RNA was isolated from COS-7 cells using Trizol (Gibco-BRL). RNA (5 μg) was hybridized overnight to ³²P-labelled antisense RNA probes (1 × 10⁵ c.p.m.) specific for actin (Ambion) and luciferase²⁷, followed by digestion with ribonucleases A and T1 (Ambion) and electrophoresis on a denaturing urea-acrylamide gel.

Purification of GST fusion proteins. BL21 bacteria were transformed with GST-PAIP-1 fusion constructs and the proteins purified on glutathione-Sepharose resin (Pharmacia).

Western blotting. The following antibodies were used: PAIP-1 antiserum (raised against the GST-PAIP-1 fusion protein, rabbit 1702) was used at a 1:1,200 dilution; monoclonal antibodies against the Flag-epitope (Kodak), the Xpress epitope (Invitrogen), and human PABP (mAb 10E10; ref. 28) were used at 1:1,000 dilution; eIF4G polyclonal antiserum (raised against the N terminus of eIF4G, rabbit 1620) was used at a dilution of 1:1,000; hybridoma supernatant of a monoclonal antibody against eIF4A (kindly provided by H. Traschel) was used at a 1:10 dilution. For immunoprecipitation, PAIP-1 antiserum was affinity-purified to give PAIP-1 and GST-specific antibodies using GST and GST-PAIP-1 coupled resins, generated with AminoLink-Plus coupling gel (Pierce).

Transient transfections. COS-7 cells were grown in 60-mm dishes (to ~50% confluence) and transfected with 0.5 μg pcDNA3-Luc and a total of 5 μg pcDNA3 plasmids (as indicated in Fig. 3a; a total of 5 μg DNA was maintained in all cases by supplementing with pcDNA3 vector) using Lipofectamine (15 μl; Gibco-BRL). Cells were collected 48 h post-transfection using luciferase lysis

buffer (Promega). Luciferase activity was measured using a BIOORBIT bioluminometer.

Received 19 November 1997; accepted 9 February 1998.

- Merrick, W. C. & Hershey, J. W. B. in *Translational Control* (eds Hershey, J. W. B., Mathews, M. B. & Sonenberg, N.) 31–69 (Cold Spring Harbor Laboratory Press, New York, 1996).
- Sonenberg, N., Rupprecht, K. M., Hecht, S. M. & Shatkin, A. J. Eukaryotic mRNA cap binding protein: purification by affinity chromatography on sepharose-coupled m7GDP. *Proc. Natl Acad. Sci. USA* **76**, 4345–4349 (1979).
- Rozen, F. et al. Bidirectional RNA helicase activity of eucaryotic translation initiation factors 4A and 4B. *Mol. Cell. Biol.* **10**, 1134–1144 (1990).
- Haghighat, A. & Sonenberg, N. eIF4G dramatically enhances the binding of eIF4E to the mRNA 5'-cap structure. *J. Biol. Chem.* **272**, 21677–21680 (1997).
- Munroe, D. & Jacobson, A. mRNA poly(A) tail, a 3' enhancer of translation initiation. *Mol. Cell. Biol.* **10**, 3441–3455 (1990).
- Jacobson, A. in *Translational Control* (eds Hershey, J. W. B., Mathews, M. B. & Sonenberg, N.) 451–480 (Cold Spring Harbor Laboratory Press, New York, 1996).
- Sachs, A. B., Sarnow, P. & Hentze, M. W. Starting at the beginning, middle, and end: translation initiation in eukaryotes. *Cell* **89**, 831–838 (1997).
- Tarun, S. J. Jr & Sachs, A. B. Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. *EMBO J.* **15**, 7168–7177 (1996).
- Tarun, S. Z., Wells, S. E., Deardorff, J. A. & Sachs, A. B. Translation initiation factor eIF4G mediates *in vitro* poly(A) tail-dependent translation. *Proc. Natl Acad. Sci. USA* **94**, 9046–9051 (1997).
- Le, H. et al. Translation initiation factors eIF-iso4G and eIF4B interact with the poly(A)-binding protein and increase its RNA binding activity. *J. Biol. Chem.* **272**, 16247–16255 (1997).
- Blanc, M. A. & Rutter, W. J. Interaction cloning: identification of a helix-loop-helix zipper protein that interacts with c-Fos. *Science* **256**, 1014–1018 (1992).
- Yan, R., Rychlik, W., Etchison, D. & Rhoads, R. E. Amino acid sequence of the human protein synthesis initiation factor eIF-4γ. *J. Biol. Chem.* **267**, 23226–23231 (1992).
- Imataka, H. & Sonenberg, N. Human eukaryotic translation initiation factor 4G possesses two separate and independent binding sites for eIF4A. *Mol. Cell. Biol.* **17**, 6940–6947 (1997).
- Lamphear, B. J., Kirchweber, R., Skern, T. & Rhoads, R. E. Mapping of functional domains in eIF4γ with picornaviral proteases. Implications for cap-dependent and cap-independent translational initiation. *J. Biol. Chem.* **270**, 21975–21983 (1995).
- Mader, S., Lee, H., Pause, A. & Sonenberg, N. The translation initiation factor eIF-4E binds to a common motif shared by the translation factor eIF-4γ and the translational repressors 4E-binding proteins. *Mol. Cell. Biol.* **15**, 4990–4997 (1995).
- Imataka, H., Olsen, H. S. & Sonenberg, N. A new translational regulator with homology to eukaryotic translation initiation factor 4G. *EMBO J.* **16**, 817–825 (1997).
- Levy-Strumpf, N., Deiss, L. P., Berissi, H. & Kimchi, A. DAP-5, a novel homolog of eukaryotic translation initiation factor 4G isolated as a putative modulator of gamma interferon-induced programmed cell death. *Mol. Cell. Biol.* **17**, 1615–1625 (1997).
- Yamanaka, S., Poksay, K. S., Arnold, K. S. & Innerarity, T. L. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. *Genes Dev.* **11**, 321–333 (1997).
- Pause, A., Methot, N., Svitkin, Y., Merrick, W. C. & Sonenberg, N. Dominant negative mutants of mammalian translation initiation factor eIF-4A define a critical role for eIF-4F in cap-dependent and cap-independent initiation of translation. *EMBO J.* **13**, 1205–1215 (1994).
- Asselbergs, F. A. M., Peters, W., van Venrooi, W. J. & Bloemendaal, H. Diminished sensitivity of re-initiation of translation to inhibition by cap analogues in reticulocyte lysates. *Eur. J. Biochem.* **88**, 483–488 (1978).
- Tarun, S. J. Jr & Sachs, A. B. A common function for mRNA 5' and 3' ends in translation initiation in yeast. *Genes Dev.* **9**, 2997–3007 (1995).
- Hentze, M. W. eIF4G: a multipurpose ribosome adaptor? *Science* **275**, 500–501 (1997).
- Gradi, A. et al. A novel functional human eukaryotic translation initiation factor 4G. *Mol. Cell. Biol.* **18**, 334–342 (1998).
- Yang, H., Duckett, C. S. & Lindsten, T. iPABP, an inducible poly(A)-binding protein detected in activated human T cells. *Mol. Cell. Biol.* **15**, 6770–6776 (1995).
- Grange, T., de Sa, C. M., Oddos, J. & Pictet, R. Human mRNA polyadenylate binding protein: evolutionary conservation of a nucleic acid binding motif. *Nucleic Acids Res.* **15**, 4771–4778 (1987).
- Haghighat, A. et al. The eIF4G-eIF4E complex is the target for direct cleavage by the rhinovirus 2A proteinase. *J. Virol.* **70**, 8444–8450 (1996).
- Donzé, O., Damay, P. & Spahr, P. F. The first and third uORFs in RSV leader RNA are efficiently translated: implications for translational regulation and viral RNA packaging. *Nucleic Acids Res.* **23**, 861–868 (1995).
- Gorlach, M., Burd, C. G. & Dreyfuss, G. The mRNA poly(A)-binding protein: localization, abundance, and RNA-binding specificity. *Exp. Cell Res.* **211**, 400–407 (1994).
- Sachs, A. B. & Buratowski, S. Common themes in translational and transcriptional regulation. *Trends Biochem. Sci.* **22**, 189–192 (1997).

Acknowledgements. We thank G. Dreyfuss and H. Traschel for monoclonal antibodies against human PABP and eIF4A, respectively; J. Bag for human PABP cDNA; M. Park for the cDNA library; Y. Svitkin for HeLa extracts; O. Donzé for luciferase plasmid; and G. Cosentino, A.-C. Gingras, B. Raught and K. Khaleghpour for comments on the manuscript. This work was supported by a grant from the MRC of Canada to N.S. A.W.B.C. was supported by a joint training award from National Health Research and Development-MRC of Canada. N.S. is a distinguished scientist of the MRC of Canada and a Howard Hughes international scholar.

Correspondence and requests for materials should be addressed to N.S. (e-mail: sonenberg@medcor.mcgill.ca). The sequence of PAIP has been deposited in GenBank under the accession number AF013758.